THE PHYSICAL SOCIETY

REPORTS ON PROGRESS IN PHYSICS

VOLUME XVII (1954)

EXECUTIVE EDITOR
A. C. STICKLAND

Price to non-Fellows £3 3*s. net* (*Postage* 2*s.*)
Price to Fellows (*one copy*) £1 7*s.* 6*d. net* (*Postage* 2*s.*)

Published by
THE PHYSICAL SOCIETY
1 Lowther Gardens, Prince Consort Road
London S.W. 7
1954

REPRINTED BY LITHOGRAPHY IN GREAT BRITAIN
BY JARROLD AND SONS, LTD, NORWICH

CONTENTS

	PAGE
Preface	iv
Nuclear Shell Structure. By M. H. L. Pryce	1
Diffraction Theory. By C. J. Bouwkamp	35
Atmospheric Electricity. By J. Alan Chalmers	101
The Physical Condition of the Solar Corona. By C. W. Allen	135
Theory of Origin of Cosmic Rays. By Edward Teller	154
Atomic Valence States and Chemical Binding. By W. Moffitt	173
Antiferromagnetism. By A. B. Lidiard	201
Horizontal Movements in the Ionosphere. By B. H. Briggs and M. Spencer	245

PREFACE

The *Reports on Progress in Physics* is under the general supervision of the Papers Committee of the Physical Society. The members of the Papers Committee were as follows:

Chairman:—Professor S. Devons.

Honorary Papers Secretary:—Dr. H. H. Hopkins.

Ordinary Members:—The President (Professor R. Whiddington), Mr. C. H. Collie, Dr. W. Ehrenberg, Mr. D W. Fry, Mr. H. G. W. Harding, Professor F. Llewellyn Jones, Professor G. O. Jones, Professor H. Jones, Dr. H. G. Kuhn, Professor H. S. W. Massey, Professor R. E. Peierls, Dr. H. Wilman.

Reports published in this Volume are being issued also as individual articles, at prices depending on their length. Information may be obtained from the Offices of the Physical Society.

A. C. STICKLAND,
Secretary–Editor.

NUCLEAR SHELL STRUCTURE

By M. H. L. PRYCE
The Clarendon Laboratory, Oxford

CONTENTS

	PAGE
§1. Introduction	2
§2. The evidence for the occurrence of nuclear shells	3
2.1. Stable nuclei	4
2.2. Abundances	6
2.3. Neutron cross sections	7
2.4. Ground-state spins of odd nuclei	7
2.5. Quadrupole moments	8
2.6. Asymmetric fission	8
2.7. Delayed neutron emitters	8
2.8. Summary of the evidence	9
§3. Models of nuclear structure	9
3.1. The individual level model	9
3.2. The one-particle model	12
3.3. Other models	13
§4. Shell structure	14
4.1. The one-nucleon levels	14
4.2. Strong spin–orbit interaction	15
4.3. The magic numbers	15
4.4. The spins of odd nuclei	17
§5. Calculation of energy levels	21
5.1. *jj*-coupling	21
5.2. Light nuclei	22
5.3. The configuration j^n	23
§6. The energy levels of even–even nuclei	24
§7. Odd–odd nuclei	26
§8. Isomerism	27
§9. Magnetic moments and quadrupole moments	28
9.1. Magnetic moments of odd nuclei	28
9.2. Magnetic moments of odd–odd nuclei	29
§10. Limitations of the individual level model—collective models	30
10.1. Configurational mixing	30
10.2. Collective models	32
10.3. The spin–orbit interaction	32
References	32

Abstract. There is convincing evidence from the study of nuclear stabilities, magnetic moments, spins, and from other sources, that there is a marked regularity of nuclear properties, particularly clearly shown by breaks at 2, 8, 20, 28, 50, 82 and 126 protons or neutrons. This is explained in terms of an individual level model of nuclear structure similar to the model used in explaining the periodicity of atomic structure. In nuclei, a strong spin–orbit interaction, of unexplained origin, plays an important part. This model leads to an excellent understanding of nuclear regularities, particularly of spins and magnetic moments. It has severe limitations, however, and, particularly when large shells are filling, it needs to be supplemented by a more 'collective' approach to nuclear structure.

§ 1. Introduction

The concept of a nuclear shell structure is as old as the hypothesis that nuclei are composed of protons and neutrons, but it is only since 1948 that it has become firmly established. It refers to a certain regularity of behaviour in the properties of nuclei as the numbers of their constituent nucleons (i.e. protons and neutrons) vary and particularly to more or less abrupt changes, notably in the binding energies, which are found to occur at certain numbers of either protons or neutrons.

There is an obvious superficial resemblance to the periodic system of the elements, in which, for example, abrupt changes in the ionization potential occur as one passes from a noble gas to an alkali metal. The interpretation of this behaviour of atoms in terms of electronic shells is well known, and the term 'nuclear shell' has come into use by analogy. In nuclei the regularities are more complex than in atoms, and it is only recently that they have been reliably established with any degree of completeness. There are several reasons for this slow development in nuclear classification. In the first place there are two different kinds of constituent particles in nuclei, protons and neutrons, of identical spin and nearly identical mass, so that nuclear properties have to be correlated with two variables, Z the number of protons, and N the number of neutrons. In atoms, on the other hand, the gross structure is entirely determined by the number of electrons Z, equal to the charge on the nucleus.

Secondly, there is nothing in nuclear physics corresponding to chemical combination, from the study of which so much information on the atomic periodic system has been derived, without resort to physical measurements. In other words, nuclear physics has suffered from the non-existence of a science of nuclear chemistry. It is only recently that a sufficiency of systematic and precise information has been accumulated to make possible the task of recognizing a genuine from a spurious pattern. In this context, the development of precise techniques, and the comparative study of many nuclei, have both played an important part.

Thirdly, it has long been believed that the successful interpretation of the periodic system of the elements in terms of the quantum theory depended on the circumstance that the motion of the electrons is dominated by the nuclear electric field and that the departure of the effects of the mutual repulsion of the electrons from a smoothed-out self-consistent field is a relatively small perturbation. The significance of this is that, to a good approximation, the electrons can be pictured as occupying individual quantum orbitals. The absence of such an overriding field inside the nucleus encouraged a belief that a similar approach would have little validity for nuclear structure. Nevertheless attempts were made, notably by Bartlett (1932), Elsasser (1933, 1934) and Guggenheimer (1934) to develop this approach, in which the germs of present ideas are to be found. The poor results obtained from calculations of nuclear binding energies by self-consistent field methods, however, discouraged further attempts at understanding nuclear structure along these lines for several years. During the intervening period, also, N. Bohr's theory of nuclear reactions and the compound nucleus, in which the emphasis is on collective motion, enjoyed success in many directions (Bohr 1936). This encouraged the view that any theory of nuclear structure must be based on notions of collective motion to the exclusion of the concept of individual motion. Indeed, it seemed difficult to reconcile Bohr's ideas with the validity of an individual approach. Recent developments, however, have shown that each approach has its merits, and that the conflict is more apparent than real.

It is the aim of the present survey to concentrate on matters of principle and on the most recent developments. Several admirable reviews of nuclear shell structure, or of certain of its aspects, have appeared in the last two years (Mayer, Moskowski and Nordheim 1951, Nordheim 1951, Rosenfeld 1951, Flowers 1952e, Goldhaber and Hill 1952, Klinkenberg 1952, Trocheris 1952, Béné 1952), and it is unnecessary to cover in detail ground already covered elsewhere.

§ 2. The Evidence for the Occurrence of Nuclear Shells

On a first study of the properties of nuclei in their ground states, the most striking regularity is the alternation in stability associated with even and odd numbers of nucleons. The word stability in this context needs a little discussion, as it is apt to be misleading. It can perhaps best be explained with reference to some such model of the nucleus as the statistical or liquid drop model, which attempts to account for the grosser features of the energy of nuclei in their ground state without going into details of structure, according to which the energy of a nucleus (including rest energy) is given by a formula (Weizsäcker 1935)

$$E = c^2(ZM_p + NM_n) - \alpha A + \beta(N-Z)^2/A + \gamma A^{2/3} + \epsilon Z^2/A^{1/3} \quad \ldots\ldots(1)$$

(M_p, M_n mass of proton and neutron, $A = N+Z$ = mass number, α, β, γ, ϵ constants). A nucleus whose energy is lower than given by the formula is said, in this connection, to be more stable than the model, and one whose energy is higher is less stable. By extension, two nuclei can be compared in stability by reference to the model. For instance, if one nucleus is more stable than the model (with the appropriate Z, N), and another is less stable than the model (again for the appropriate, but different, Z and N), then the first is said to be the more stable of the two. Stability in the present sense should not be confused with stability against β-decay.

It is found that nuclei with even Z and even N (even–even nuclei) are always more stable than nuclei with odd Z and even N, or even Z and odd N (odd–even and even–odd nuclei), which in turn are more stable than odd–odd nuclei (with the exception of the self-conjugate odd–odd nuclei ^{2}H, ^{6}Li, ^{10}B, ^{14}N). This simple rule is often incorporated into the semi-empirical formula (1) by adding a term $-\delta, 0$ or $+\delta$ for even–even, even–odd and odd–even, or odd–odd nuclei. The 'pairing energy', δ, is not a constant, but varies very roughly as $A^{-1/2}$ (Hitchcock 1953). This alternation of stability compared with the model nucleus, combined with the requirements for stability against β-decay, in the naturally occurring light nuclei ($A<40$) gives rise to periodic behaviour every time two protons and two neutrons are added. As the α-particle, ^{4}He, consists of two protons and two neutrons, such behaviour would be expected if α-particles acted as subsidiary units in nuclear structure. It has been explained as a natural consequence of the symmetry characteristics of the wave function of the nucleus imposed by the Pauli principle, taken in conjunction with quite general and simple hypotheses concerning nuclear forces (Wigner 1937a). It also fits naturally into the theories of shell structure about to be discussed. It should not be taken as evidence for the existence of α-particles as well-defined units in the nucleus (α-particle model).

The deeper seated shell structure is masked by this α-particle periodicity in light nuclei, and is more clearly revealed only in heavier nuclei. Its most obvious manifestations are connected with a discontinuity in the general trend of stability at certain definite nucleon numbers, the so-called magic numbers,

corresponding to closure of a shell. Nuclei that possess a magic number of either protons or neutrons have a greater stability (in the sense defined above) than others in their immediate neighbourhood. The stability tends to increase gradually as the nucleon number increases towards a closed shell, and then to decrease rather abruptly beyond the closed shell. Other nuclear properties also are correlated with the shell structure, notably the ground state spins of odd nuclei, their magnetic moments and electric quadrupole moments.

In spite of the obscuring effects of the α-particle periodicity, a straightforward comparison of real nuclei with the standard model shows a marked break at ${}^{4}_{2}\mathrm{He}$ and again at ${}^{16}_{8}\mathrm{O}$, indicating the existence of the magic numbers 2 and 8, as first pointed out by Elsasser (1933). Guggenheimer (1934), studying the relation between Z and N in naturally occurring nuclei (essentially this means those which are stable against β-decay) inferred the existence of other magic numbers. His list (2, 8, 18, 32, 50, 60, 126) contains numbers which nowadays are considered to be spurious, or only 'semi-magic', and omits others now known to be 'magic'. Mayer (1948), using more up-to-date information, corrected and extended the list. Closed shells are nowadays recognized as occurring at the following magic numbers:

A 2, 8, 20, (40), ..., $(=n(n+1)(n+2)/3)$

B 2, (6), (14), (28), 50, 82, 126. $(=n(n^2+5)/3)$

Those included in parentheses do not show such marked breaks in stability as the others, and are sometimes considered as only semi-magic numbers.

It will be noticed that the magic numbers have been arranged in two sequences, A and B, each of which can be represented by a simple numerical formula, namely $\mathrm{A}=n(n+1)(n+2)/3$, $\mathrm{B}=n(n^2+5)/3$. The A-sequence shows strong closed shell effects in light nuclei, while for heavier nuclei the closed shells belong to the B-sequence.*

The evidence on which the existence of these closed shells is deduced will now be summarized under the appropriate headings.

2.1. *Stable Nuclei*

For a given mass number A, only a limited number of the possible nuclei exist as a stable nucleus, because of the possibility of β-decay. With only a very few exceptions, if A is odd, there is only one stable isobar. If the semi-empirical formula (1) held exactly, the stable isobar would be determined by minimizing with respect to Z (or N) for fixed A, and taking the nearest integer. A plot of Z against A, or some equivalent plot, for odd stable nuclei, would then be expected to lie close to a smooth curve. As, however, there is a shell structure, the actual plot shows kinks (Kohman 1948, 1952, Wapstra 1952, Coryell 1953, Bouchez, Robert and Tobailem 1953, Green and Edwards 1953). Even values of A, because of the alternation of stability between even–even and odd–odd nuclei, can have several stable isobars, and a study of the number of stable even isobars with increasing A provides further evidence of discontinuities.

Plots of the number of stable isotopes for a given Z, against Z, are another way of revealing this information. An example may be quoted. Tin ($Z=50$) is unique in having as many as ten stable isotopes. This is interpreted as due to the

* The number 2 has been included in both sequences.

great stability of nuclei containing 50 protons. A plot of the number of stable isotones (nuclei with an equal number of neutrons) against N shows a similar behaviour. The two largest numbers of isotones are for $N=82$ (7 isotones) and $N=50$ (6 isotones).

This method of analysis can be refined by using β-decay data to trace out the actual variation in nuclear energy with Z for a given mass number, instead of merely plotting the nearest integer to the minimum (Coryell 1953). In this way the non-integral value Z_A which minimizes the energy can be plotted and gives a smoother curve in which the breaks show up more clearly. Naturally, the more refined is the method of analysis, the greater is the number of minor fluctuations which are revealed, so that it is sometimes difficult to decide whether a number should be classed as a semi-magic number or not. The numbers 50, 82, 126 stand out clearly as shell closures.

Nuclear masses and radioactivity.

If nuclear masses could be determined very accurately, a straightforward comparison with the standard model nucleus (1) would be a good way of studying the periodicity of stability. Unfortunately only the masses of relatively light nuclei are known with the desired accuracy, and a direct attack along these lines is subject to considerable errors. Modern techniques of mass spectrometry, however, are rapidly advancing to the stage when this will be practicable. Even so, the plot of binding energy per nucleon, reproduced in fig. 1 from Duckworth

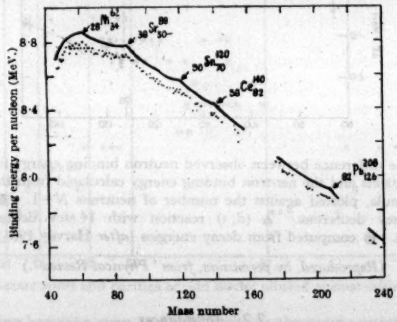


Fig. 1. Binding energies per nucleon (*after* Duckworth 1952).
(*Reproduced, by permission, from* '*Nature*'.)

(1952), is revealing. At present, for the medium and heavy nuclei, only mass differences of neighbouring nuclei are known with adequate precision, through measurements of energy released in radioactive disintegrations and nuclear reactions. Several authors have constructed 'energy surfaces' in this way, and analysed them in terms of shell structure (Wapstra 1951, 1952, Green and Edwards 1953). Harvey (1951) has studied neutron binding energies and compared them with those calculated according to the semi-empirical formula. The result

(fig. 2) shows striking discontinuities at $N=50$, 82, 126. An extensive empirical analysis of available data is given by Feather (1953), who gives tables of neutron and proton binding energies. These have been derived using all the available information from nuclear reactions and radioactive decay. Similar tables have been given by Way and Wood (1951). An analysis of the disintegration energies of heavy α-emitters has been given by Glueckauf (1948), Perlman, Ghiorso and Seaborg (1950), Pryce (1950). Pronounced discontinuities occur at $Z=82$ and $N=126$. The occurrence of α-radioactivity in the rare earth region also points to a similar phenomenon at $N=82$.

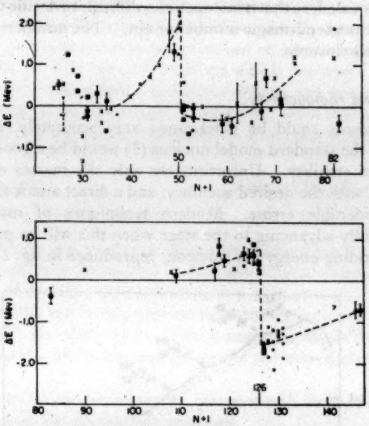


Fig. 2. ΔE is the difference between observed neutron binding energy for a nucleus with $N+1$ neutrons and the neutron binding energy calculated from the semi-empirical mass formula, plotted against the number of neutrons $N+1$. ■ (d, p) reactions with 14 MeV deuterons. ▲ (d, t) reaction with 14 MeV deuterons. × (γ, n) thresholds. ● computed from decay energies (*after* Harvey 1951).

(*Reproduced, by permission, from* ' *Physical Review*'.)

2.2. *Abundances*

The relative abundance of the various nuclear species shows a marked correlation with magic numbers. Although there is still disagreement on the origin of the elements, all serious theories agree that the more relatively stable a nucleus is, the greater is the chance of its having a high abundance, so that abundance can be regarded as connected with stability. Nuclei with 8, 20, 50, 82, 126 protons or neutrons tend to be particularly abundant. Examples are ^{16}O ($Z=N=8$), ^{40}Ca ($Z=N=20$), ^{120}Sn ($Z=50$), ^{88}Sr, ^{89}Y, ^{90}Zr ($N=50$), ^{138}Ba, ^{139}La, ^{140}Ce ($N=82$), ^{208}Pb ($Z=82$, $N=126$). It has been suggested that the very high abundance of $^{56}_{26}Fe$ is connected with the high stability of ^{56}Ni, with 28 protons and 28 neutrons, which after formation decayed in two steps to ^{56}Fe.

2.3. *Neutron Cross Sections*

The capture cross sections for medium energy neutrons are abnormally low for nuclei with $N = 50$, 82, 126 (von Halban and Kowarski 1938, Griffiths 1939, Mescheryakov 1945, Hughes, Spatz and Goldstein 1949). The capture cross section for thermal neutrons is known to behave capriciously because of the importance of narrow resonance levels occurring near the dissociation energy of the final nucleus. Nevertheless, as emphasized by Flowers (1952e), they show remarkable minima near $N = 20$, 50, 82 and 126.

2.4. *Ground-State Spins of Odd Nuclei*

As will be discussed later, an interpretation of nuclear spins and magnetic moments in terms of the one-particle model, or less crudely in terms of the individual-particle model, enables one to assign an orbital angular momentum to the state of the odd nucleon in an odd nucleus. Thus, for example, a nucleus with spin $I = 5/2$ can have orbital momentum either 2 ($d_{5/2}$) or 3 ($f_{5/2}$), its magnetic moment being considerably different in the two cases. Figure 3 shows a plot of observed spins, interpreted in this way, against the odd-nucleon number. For reasons which will be apparent later, the orbits have been arranged in two sequences according to their parity, $s_{1/2}$, $d_{3/2}$, $d_{5/2}$, $g_{7/2}$, $g_{9/2}$, $i_{11/2}$, ... (even parity, denoted 1/2+, etc.), and $p_{1/2}$, $p_{3/2}$, $f_{5/2}$, $f_{7/2}$, $h_{9/2}$, $h_{11/2}$, ... (odd parity, denoted 1/2−, etc.). Four points on this plot appear in anomalous positions and have been marked as open circles to distinguish them. Otherwise the diagram shows a markedly regular pattern, the

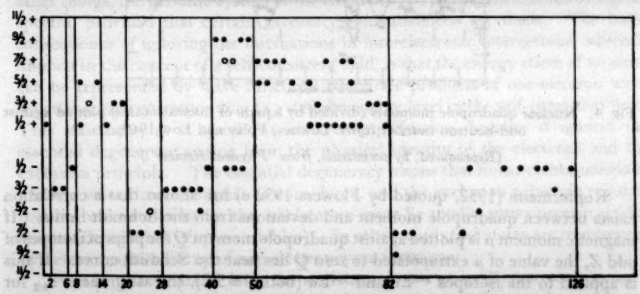


Fig. 3. Ground-state spins and parities of odd nuclei plotted against odd-nucleon number.

points alternating between even and odd parity at the magic numbers 2, 8, 20, 40 and 82, and showing other types of jump at 6, 14, 28 and 50. Where different nuclei with the same odd nucleon number have different spins, more than one point is shown. It is remarkable, however, how often the nucleon number determines a unique spin. The interpretation of these features, and of the anomalous points, is discussed later. This type of diagram was first used by Latham (1947).

Béné (1952) has analysed magnetic moments by introducing a parameter which takes the value 0 when the magnetic moment corresponds to the Schmidt value for spin and orbit antiparallel and the value 1 for spin and orbit parallel. A plot of this parameter against Z for odd–even nuclei shows discontinuities at $Z = 8$, 20, 40 and 82, and anomalies associated with $N = 28$ and 40.

2.5. *Quadrupole Moments*

Nuclear quadrupole moments can be either positive or negative. There is a definite correlation with shell structure, nuclei just above a closed shell having negative quadrupole moments, and those just below positive. There is a reasonably regular variation, rising to a maximum for a roughly two-thirds full shell (Gordy 1949, Hill 1949, Townes, Foley and Low 1949, Rosenfeld 1951). Proton and neutron shells are both effective, indicating some kind of collective or co-operative motion of protons and neutrons. A plot of quadrupole moments against odd nucleon number is shown in fig. 4. We return to this subject later.

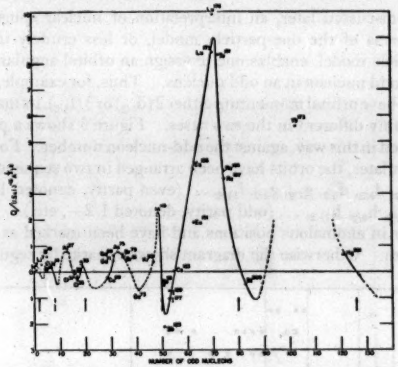


Fig. 4. Nuclear quadrupole moments (divided by square of nuclear radius) plotted against odd-neutron number (*after* Townes, Foley and Low 1949).

(*Reproduced, by permission, from* '*Physical Review*'.)

Kopfermann (1952, quoted by Flowers 1952 e) has shown that a correlation exists between quadrupole moment and deviations from the Schmidt limits. If magnetic moment μ is plotted against quadrupole moment Q for pairs of isotopes of odd Z, the value of μ extrapolated to zero Q lies near the Schmidt curve. If this is applied to the isotopes ^{151}Eu and ^{153}Eu (both $I=5/2$), the assignment $d_{5/2}$ for both is supported, whereas on the basis of μ alone ^{153}Eu would appear to be $f_{5/2}$, in violation of the regular pattern of fig. 3.

2.6. *Asymmetric Fission*

The effects of the enhanced stability of fission products in the vicinity of $N=50$ on the one hand and $N=82$ on the other hand has been discussed by several authors as a possible explanation of asymmetric fission (Mayer 1948, Kowarski 1950, Meitner 1952, Hill and Wheeler 1953). Hill and Wheeler are of the opinion that this explanation is incorrect.

2.7. *Delayed Neutron Emitters*

The fission products ^{87}Br ($N=52$) and ^{137}I ($N=84$) emit delayed neutrons. The process is one first of β-emission (N decreases by one unit) to a highly excited

state of the daughter nucleus, which then boils off a neutron. Its occurrence indicates the relatively loose binding of the 51st and 83rd neutron respectively, as is expected from shell closure at $N=50$ and 82.

2.8. *Summary of the Evidence*

From the foregoing it can be deduced that effects which can be associated with shell closure occur for nucleon numbers 2, 8, 20, 28, 40, 50, 82 and 126. This is true independently for protons and neutrons, a striking fact which must be considered as significant in any theory of nuclear structure.

Quantitative estimates of the discontinuity in binding energy of the last nucleon (referred to the standard nucleus) are not very easy to make, as they are complicated by the even–odd alternation and the uncertainties of extrapolating somewhat uncertain data. They have been reviewed by Coryell (1953), who concludes that the jump is between 1 and 2 MeV at $Z=20$, 28, 50, 82 and $N=20$, 28, 50, 82, 126, while no significant jump is associated with N or $Z=40$.

§ 3. Models of Nuclear Structure

3.1. *The Individual Level Model*

It has already been mentioned that in atomic structure a representation in terms of an approximation in which the interactions between the electrons are replaced by a self-consistent field is remarkably successful in explaining, among other things, the periodic system of the elements. The success extends to atomic spectra, provided that certain necessary generalizations are made. The basic consequence of ignoring the fluctuations in interelectronic interactions, which is implicit in the concept of a self-consistent field, is that the energy states of an atom can be represented by wave functions which are products of one-electron wave functions, each corresponding to a definite energy level in the self-consistent field. This statement, as it stands, is an oversimplification, because it ignores the essential degeneracy arising from the physical identity of the electrons and the exclusion principle. The essential degeneracy means that linear combinations of product wave functions must be considered, and the exclusion principle requires that the wave function for the whole atom must be totally antisymmetrical (due account being taken of spin variables). In other words, the states are represented by determinants built up of one-electron wave functions.

Even in this modified form the statement is not generally correct because a further degeneracy remains, arising from the degeneracy of the one-electron states, due to the spherically symmetrical nature of the field of force (states with different orientations of the same angular momentum having the same energy). So the most general expression for the approximate wave function is a linear combination of determinants. If the mutual interaction of the electrons were completely represented by the self-consistent field any linear combination would be an energy state (eigenstate), but as it is not, it is necessary to take some account of the actual interactions. The result is to split up the states into several energy levels. The group of states which all have the same energy before the detailed interaction is considered is called a configuration. A configuration is specified by the one-electron levels in the determinantal wave functions (e.g. $1s^2\,2p$ means two electrons in the 1s-state, i.e. one in each spin sub-state, and one in the 2p state). The groups of states which have equal energy after interaction is considered are called states or

levels.* Only for a closed shell configuration does the wave function reduce to a single determinant, corresponding to a single level.

That this is an approximation is to be emphasized. How good an approximation it is depends on how much the mutual interaction differs from a self-consistent field. Another important factor is whether there are other configurations with nearly the same energy, for then the real wave function may contain appreciable amounts of several configurations (perturbation, or configuration interaction). In any case the real wave function must contain a certain amount of admixture from other configurations (an infinite number in fact), but under favourable circumstances these admixtures will be small.

We wish to apply a similar approximation in the study of nuclear structure. It must be recognized at the outset, though, that the conditions for its being a good approximation are much less favourable than with atoms. The fact that the forces between nucleons are of short range has the consequence that in a nucleus any nucleon is effectively in strong interaction with only its immediate neighbours, so that fluctuations away from a self-consistent field are probably more important than in the atomic problem. Again there is nothing inside the nucleus analogous to the nucleus inside the atom. In the atomic problem the nucleus is so much heavier than the electrons that it can be treated as a stationary point, about which the self-consistent field is centred. In the nuclear problem the self-consistent field must be considered as centred relative to the centre of mass of the nucleus, so that the smoothed-out potential is not a function of the coordinates of one nucleon only. This has the consequence that the wave function, even in the lowest approximation of an effective field, cannot strictly be written as a product or determinant of single-nucleon wave functions. In applications of the method to complex nuclei this feature is ignored, the self-consistent field being considered as centred about some unspecified point, with consequent loss of accuracy.

At present too little is known about the details of nuclear forces for any precise calculation of the form of the effective field to be made. As it turns out this is not so very important. Quite qualitative statements are enough for considerable progress to be possible. There is independent evidence that inside the nucleus the density of nucleons is rather constant, and since the range of nuclear forces is known to be short (comparable with the linear dimensions of small nuclei), the effective potential can be expected to be fairly constant inside the nucleus, falling off rather sharply at the nuclear 'radius'. This can be described as a 'square well with rounded edges'.

Different potentials must be taken for protons and neutrons since the former are influenced by the electrostatic (coulomb) field and the latter are not. For light nuclei the coulomb field is small compared with the nuclear forces, but for heavy nuclei it can become important. It then has the effect of raising the centre of the potential well for protons relative to the peripheral region (wine-bottle potential). Figure 5 shows the type of potential usually assumed. The depth of the potential well is about 25 to 30 MeV.

Available evidence seems to be consistent with the hypothesis that the specifically nuclear forces are the same between any pair of nucleons, whether these be two protons, two neutrons or a proton and a neutron. This circumstance leads to a formal simplification of the problem. It is possible to describe

* As the purpose here is not to give an account of atomic structure, but of the principles underlying the theory which are of use in nuclear structure, no reference is made to terms, of which the levels form the fine structure.

the proton and the neutron as two different states of the same kind of particle, the exclusion principle now requiring complete antisymmetry of wave function with respect to all variables, including the new two valued variable specifying whether a nucleon is a proton or a neutron, called the isobaric spin. The wave function can now be written as a determinant, or linear combination of determinants, of individual states for each nucleon. Such a representation may appear more complicated than the simple proton–neutron representation, since it leaves unspecified whether a given nucleon is a proton or neutron, but it leads to the same physical results as the more elementary formulation. In problems where the coulomb energy is relatively unimportant, as in light nuclei, it has the great advantage that use can be made of group theory in classifying the energy levels (Wigner 1937 a, Jahn 1950, Flowers 1952 b). No explicit use of the isobaric spin formalism will be made in this Report beyond referring to certain wave functions as determinants without distinction between protons and neutrons.

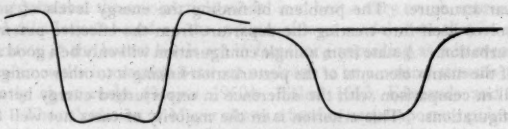


Fig. 5. Proton and neutron wells.

If we now adopt such an effective potential for our starting approximation, we can regard the energy levels of the nucleus as represented by determinants of one-nucleon wave functions, each of which corresponds to a one-nucleon energy level. In a nucleus with Z protons and N neutrons we specify Z proton-states and N neutron-states as occupied. Owing to the degeneracy of the individual states, however, this specification is not usually enough. To enable us to talk more clearly about nuclear states we will draw a distinction between a group of one-nucleon states all with the same energy, which we shall call a level, and the separate states of that level, which we shall call sub-levels. We then introduce the idea of a configuration. We specify how many nucleons there are in each level. Because of the exclusion principle, not more than one nucleon can go in any sub-level. All the states consistent with this specification together form a configuration. In this order of approximation (which we shall call the zero order) all the states of one configuration have the same energy.

In general there is more than one state belonging to a given configuration. An exception is when each one-nucleon level is either completely full or completely empty. We then speak of a closed sub-shell configuration or, in certain cases, a closed shell configuration.

In the general case the departure of the true interaction between the nucleons from the effective potential well results in a splitting of the configuration into a number of different energy levels of the nucleus. This poses a problem of considerable complexity to which we shall return later. At this stage we merely remark that to each level there corresponds a definite angular momentum, or nuclear spin, I.

We are usually interested in the ground state of a nucleus or some fairly low lying energy level, so that a study of the lowest configuration has a particular importance. This configuration is obtained by filling the lowest available

one-nucleon levels. In this way all the lowest proton levels up to a certain point are full, the remaining protons going into the next highest level, and similarly for the neutrons. The partial configuration consisting of the full levels is usually referred to as the core. The core configuration is non-degenerate, and the degeneracy of the full nuclear configuration is then just that of the partial configuration made up of the 'loose' protons and neutrons in the two unfilled shells. Even with this simplification the problem of classifying the energy levels can be formidable.

A model of nuclear structure in which the description of the energy levels in terms of a single configuration is considered a good working approximation will be called an individual level model. It is also often referred to as an individual particle or independent particle model.

It is clear that the most general nuclear state can be regarded as made up of a superposition of states from all possible configurations, so that in principle the individual level model can be used as the starting point for a rigorous theory of nuclear structure. The problem of finding the energy levels of a nucleus then resolves itself into treating the departure from the effective potential well as a perturbation. A state from a single configuration will only be a good approximation if the matrix elements of the perturbation linking it to other configurations are small in comparison with the difference in unperturbed energy between the two configurations. This criterion is in the majority of cases not well fulfilled, from which one deduces that the individual level model is often only a poor approximation.

3.2. *The One-Particle Model*

It is a striking experimental fact that the spin of all even–even nuclei is zero (in the ground state). No exception is known. It follows that they have zero magnetic moment. Schmidt (1937) has pointed out that the magnetic moments of odd nuclei can be understood if one assumes that only the unpaired nucleon contributes to the spin and magnetic moment of the nucleus. On this picture an even number of like nucleons is to be regarded as always pairing off to give zero resultant angular momentum and magnetic moment, undisturbed by the presence of an unpaired particle. The spin I of an odd nucleus is then equal to the angular momentum j of the unpaired particle, made up of orbital angular momentum l and spin angular momentum $\frac{1}{2}$, which may be either parallel ($j=l+\frac{1}{2}$) or antiparallel ($j=l-\frac{1}{2}$). The magnetic moment associated with the unpaired particle is easily calculated, on the assumption that it arises from the addition of an orbital and a spin magnetic moment. When plotted against j it lies on one of two curves, according as spin and orbit are parallel or antiparallel, called the Schmidt curves or Schmidt limits.

Actually the magnetic moments of odd nuclei do not lie exactly on the Schmidt limits, but scatter rather widely between them. With few exceptions, however, the experimental points fall into two reasonably distinct groups near one or other of the Schmidt limits. This indicates that the foregoing, so-called one-particle, model, though rough, contains some measure of truth. An immediate consequence is that a determination of spin and magnetic moment enables one to decide whether spin and orbit are parallel or antiparallel, and so to fix l ($=j\mp\frac{1}{2}$ respectively), and hence the parity of the state of motion of the odd particle.*

* A state is of even parity (+) if l is even, and odd (−) if l is odd.

The notion of the parity of a nuclear state is an important one, and plays a considerable part in discussions of radioactive and radiative transitions. It is implicit in the one-particle model that the paired-off nucleons are in a state of even parity, and consequently that the parity of the nucleus as a whole is determined by the parity of the odd nucleon.

The one-particle model may be taken as an empirical rule determining the spin of the ground state of a nuclear configuration. If outside the core there are n protons in a level j and n' neutrons in a level j', it states that in the lowest nuclear level the resultant angular moment of j^n is 0 or j, according as n is even or odd, and similarly for $j'^{n'}$. If both n and n' are odd, it makes no definite prediction beyond stating that I takes one of the integral values between $|j-j'|$ and $j+j'$, which are the possible vector resultants of j and j'. This rule is found to give good agreement with observed magnetic moments of odd–odd nuclei in most of those cases where a reasonable assignment to j and j' can be made.

The considerable deviation of most magnetic moments from the Schmidt limits suggests that the one-particle model is at best a very rough approximation. Detailed calculations based on the individual level model show that in a nuclear configuration j^n the lowest levels are in fact 0 or j according as n is even or odd, if one assumes the interaction between nucleons to be a central attractive force of sufficiently short range (Mayer 1950b, Talmi 1952, Edmonds and Flowers 1952b). The calculation of Mayer shows a further very important point, namely that the binding energy of an added nucleon alternates with odd and even n. In the limit of very short range forces the binding energy of an odd nucleon, referred to the unperturbed configuration, is zero, while that of an even nucleon is considerable. The theoretical expression for this 'pairing energy' contains a factor $2j+1$ outside an expression which probably does not vary over very wide limits for nuclei of comparable size, thus indicating that for one-nucleon levels of high j it can be very appreciable. This fact has a bearing on shell structure (§4.4).

These calculations also reveal that it is not correct to suppose that in a configuration containing an odd number n of nucleons the first $n-1$ nucleons pair off to a zero resultant angular momentum: but it is true that the dominant part of the wave function comes from such an arrangement. This has an important bearing on the deviations of the magnetic moments from the Schmidt limits in the nuclei containing a mixed configuration of protons and neutrons outside a core.

The calculations further show, again provided the range of the forces is sufficiently short, that there is a considerable gap in energy between the lowest level and the rest, which fall together in groups, and this suggests the following useful empirical refinement of the one-particle model. In assessing the relative positions of the lowest levels of several close configurations one ascribes an unperturbed energy to the configuration together with as many times the (negative) pairing energy as there are completed pairs of like nucleons in a given one-nucleon level. This is a very rough, but rather useful, rule.

3.3. *Other Models*

We shall not discuss the α-particle model of nuclear structure (Wefelmeier 1937, Hafstad and Teller 1938) beyond mentioning that it appears to give, within limits, a good account of the ground and low excited states of the light α-nuclei (i.e. $N=Z=$even). These features are also explainable on the individual level

model, and it is interesting to note how the same conclusions can be reached from such dissimilar starting points.

Where the individual level model indicates that many configurations lie close together at the bottom, a simple representation in terms of a single configuration becomes impossible and a collective type of motion is to be expected. It has been suggested that then a model in terms of a distorted core should be used (Rainwater 1950, A. Bohr 1951, 1952). Collective motion will be discussed in a later section.

§4. Shell Structure

4.1. *The One-Nucleon Levels*

We shall now adopt an individual level model of nuclear states as our provisional point of departure. We must now consider the ordering of the different energy levels in the potential well. At first sight it would seem that as we know so little about the actual shape of the well we could not expect to make much progress. Actually, however, the order in which the levels occur is very insensitive to well shape, and almost any reasonable hypothesis gives much the same results.

It is convenient to start by discussing the levels in a three-dimensional harmonic oscillator (parabolic well), because they reproduce the major features of the classification in any nuclear well. It should first be observed that the square well with rounded edges is in some ways intermediate between a square well and a harmonic potential, so that we may reasonably expect the nucleon levels to be intermediate between the levels in a square well and those of a harmonic oscillator. And since the levels in a square well show some resemblance with those of a harmonic oscillator we have some reason for proceeding in this way.

The lowest level of a (spinless) harmonic oscillator is non-degenerate, and has angular momentum $l=0$. It will be denoted by the spectroscopic symbol 1s. The second level is triply degenerate, with $l=1$, and will be denoted 1p. The third has sixfold degeneracy and corresponds to the two angular momenta $l=2,0$. Any departure from the parabolic shape of the well resolves the degeneracy of the two values of l. The levels will be called 1d and 2s.* Going on in this way one finds degenerate levels 1f and 2p ($l=3,1$), then 1g, 2d and 3s ($l=4,2,0$), 1h, 2f, 3p ($l=5,3,1$), 1i, 2g, 3d, 4s ($l=6,4,2,0$) and so on. In a deep square well the sequence of levels is 1s, 1p, 1d, 2s, 1f, 2p, 1g, 2d, 1h, 3s, ... The degeneracy of the harmonic oscillator is broken down, but the levels up to 2p remain in the same groupings (fig. 6). Beyond this point the order of the levels becomes dependent on the radius of the potential well in relation to its depth. The rounding off of the potential well tends to bring the levels nearer to the grouping of the oscillator. The degenerate groups of levels of the oscillator alternate in parity (1s: +, 1p: −, 1d, 2s: +, 1f, 2p: −, etc.), and this feature remains characteristic.

It is well to stress at this point that the radius of the potential well increases steadily with increasing mass number of the nucleus, and that this fact has important consequences. In a well of finite depth and radius only a finite number of bound levels exists, as contrasted with the infinite number for a harmonic oscillator. As the radius increases so does the number of bound

* The spectroscopic notation varies with different authors. Some call the levels 1s; 2p; 3d, 2s; 4f, 3p; 5g, 4d, 3s; etc. The notation used here is that of Mayer.

states. For example, for $A=5$, although the assumption of anything like a potential well is perhaps too crude, it would appear that only the 1s state is bound, whereas for $A=200$, with the usual assumptions about well depth and radius, the states up to about 1k, 2h, 3f are bound.

4.2. *Strong Spin–Orbit Interaction*

The previous discussion has been confined to spinless particles. Nucleons, however, possess a spin angular momentum of $\frac{1}{2}$, with a consequent doubling of the number of states available to them. Although our present knowledge of nuclear forces gives no reason to suppose that there is any very strong interaction between spin orientation and orbital motion, it has been suggested independently by Haxel, Jensen and Suess (1948, 1949, 1950) and by Mayer (1949, 1950a) as a hypothesis to explain the magic numbers that there is such an interaction. The effect will be to split a level of given l into two levels characterized by the total angular momentum $j=l\pm\frac{1}{2}$. It is assumed that the level with spin and orbit parallel ($j=l+\frac{1}{2}$) is lower than the other. An exception of course is an s-level, for which the only resultant possible is $j=\frac{1}{2}$.

Although nothing is known *a priori* about the mechanism of spin–orbit interaction, it is plausible to believe that the splitting will be greater for large l than for small l. In this way it is possible for the lowest level of an oscillator group, which in the square well is always the one of that group with the highest l, to be split in such a way that the $j=l+\frac{1}{2}$ level comes down into the previous group. This is the key hypothesis which leads to an explanation of the observed magic numbers of the B-sequence, namely 28, 50, 82 and 126.

4.3. *The Magic Numbers*

If it is assumed that the one-nucleon levels are grouped together in much the same way as in the oscillator potential, and that spin–orbit interaction is unimportant, then it is easy to see that well-marked closed shells will occur at nucleon numbers 2, 8, 20, 40, 72, etc. (the A-sequence). If, on the other hand, it is assumed that spin–orbit interaction depresses the lowest level into the group below, as just discussed, then the well-marked closed shells will occur at 2, 6, 14, 28, 50, 82, 126, etc. (the B-sequence).

This is illustrated in table 1, adapted from Mayer (1949) and in fig. 6, taken from Haxel, Jensen and Suess (1950).

The lowest state, $1s_{1/2}$, holds two nucleons. There is thus a closed shell at 2. In the next state the spin–orbit interaction is not yet strong enough to separate the $1p_{3/2}$ and $1p_{1/2}$ levels by an amount comparable with well spacings, so the next closed shell occurs when all the 6 p-states are occupied, at $2+6=8$.

Continuing, the next states, $1d_{5/2}$, $1d_{3/2}$, $2s_{1/2}$, are sufficiently close to constitute a single shell. They number 12 in all. The next shell therefore closes at $8+12=20$. Proceeding in this way one obtains the A-sequence of magic numbers.

At this point the strong spin–orbit interaction becomes effective, and the $1f_{7/2}$ level (8 sub-levels) is depressed to a position intermediate between the previous shell and the other levels of the group ($1f_{5/2}$, $2p_{3/2}$, $2p_{1/2}$), forming a subshell which is sufficiently separated in energy from the others to rank as a shell. This shell closes at $20+8=28$.

The next shell comprises the $1f_{5/2}$, $2p_{3/2}$, $2p_{1/2}$ and $1g_{9/2}$, the last coming down, because of spin–orbit interaction, from the next higher group of levels. As it contains 22 sub-levels this shell closes at $28+22=50$.

In a similar manner the next shell, made up of the 32 sub-levels of $1g_{7/2}$, $2d_{5/2}$, $2d_{3/2}$, $3s_{1/2}$ and $1h_{11/2}$, closes at $50+32=82$. And the one after that, with the 44 sub-levels of $1h_{9/2}$, $2f_{7/2}$, $2f_{5/2}$, $3p_{3/2}$, $3p_{1/2}$ and $1i_{13/2}$, at $82+44=126$. This is the last observed magic number. If the pattern continued in the same way,

Table 1

Oscillator number	Square well	Spin term	Number of states	Number in shell	Total number
	4s	$4s_{1/2}$			
	3d	$3d_{3/2}$			
6		$3d_{5/2}$	6		
	2g	$2g_{7/2}$	8		
		$2g_{9/2}$	10		
	1i	$1i_{11/2}$	12		126
		$1i_{13/2}$	14		
	3p	$3p_{1/2}$	2		
		$3p_{3/2}$	4	44	
5	2f	$2f_{5/2}$	6		
		$2f_{7/2}$	8		
	1h	$1h_{9/2}$	10		82
		$1h_{11/2}$	12		
	3s	$3s_{1/2}$	2		
4	2d	$2d_{3/2}$	4	32	
		$2d_{5/2}$	6		
	1g	$1g_{7/2}$	8		50
		$1g_{9/2}$	10		
	2p	$2p_{1/2}$	2	22	
3		$2p_{3/2}$	4		
	1f	$1f_{5/2}$	6		28
		$1f_{7/2}$	8	8	
	2s	$2s_{1/2}$	2		20
2	1d	$1d_{3/2}$	4	12	
		$1d_{5/2}$	6		8
1	1p	$1p_{1/2}$	2	6	
		$1p_{3/2}$	4		2
0	1s	$1s_{1/2}$	2	2	

the next shell would contain $1i_{11/2}$, $2g_{9/2}$, $2g_{7/2}$, $3d_{5/2}$, $3d_{3/2}$, $4s_{1/2}$ and $1k_{15/2}$ (58 sub-levels), closing at $126+58=184$, which is far outside the range of stable nuclei (the nuclei with the largest number of neutrons at present known are ^{243}Pu, ^{244}Am and ^{245}Cm, with 149 neutrons).

Although protons and neutrons move in different effective fields, the difference is insufficient to change the above grouping of levels into shells. Nevertheless it is to be expected that the order in which protons and neutrons fill the levels (sub-shells) is different for the higher shells, and this is indeed found. Qualitatively it is to be expected that in the wine-bottle potential applicable to

protons, states of high angular momentum l are depressed relative to states of lower l, compared with neutrons. An illustration of this is furnished by the first state occupied after closure of the 82-shell. The 83rd proton occupies the $1h_{9/2}$ level, but the 83rd neutron occupies the $2f_{7/2}$ level.

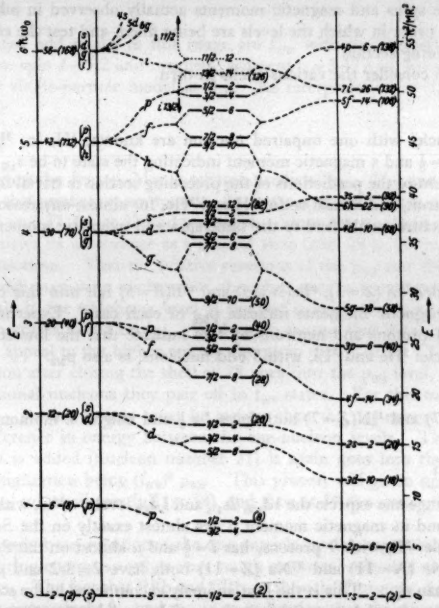


Fig. 6. Energy levels of a nucleon (*after* Haxel, Jensen and Suess 1950). Left, oscillator potential. Right, square well. Intermediate are levels, with spin–orbit interaction. (Notation of well-levels different from text.)

(*Reproduced, by permission, from* '*Zeitschrift für Physik*'.)

Other explanations of the magic numbers based on potential wells of special shapes, without strong spin–orbit interactions, have been proposed (Feenberg and Hammack 1949, Nordheim 1949), but they involve implausible assumptions and at the present time are only of historical interest.

4.4. *The Spins of Odd Nuclei*

As the order of levels in the potential is not very sensitive to its radius a discussion of which levels are completely filled and which are incompletely filled in a particular nucleus can proceed without paying much attention to the fact that as A increases the well radius also increases. For the present we shall adopt the single particle model.

In an odd nucleus there is an odd number of one kind of nucleon and an even number of the other kind. The latter are considered to be paired off, giving no contribution to the spin and magnetic moment. Of the former, all but one are considered to be also paired off, leaving the odd nucleon to determine the spin and magnetic moment. If the one-particle model is trustworthy, therefore, a study of the spins and magnetic moments actually observed in odd nuclei will indicate the order in which the levels are being filled, and test the considerations of the preceding section.

We shall consider the various shells in turn.

$0<Z, N<2$

Four nuclei with one unpaired nucleon are known, ^{1}H, ^{1}n, ^{3}H, ^{3}He. All four have $I=\frac{1}{2}$ and a magnetic moment indicating the state to be $s_{1/2}$ and not $p_{1/2}$. The fulfilment of the predictions of the preceding section is trivial for the proton and the neutron, and almost so for ^{3}H and ^{3}He, for almost any reasonable model of their structure would lead to the same spin and magnetic moment.

$2<Z, N<6$

The nuclei ^{7}Li ($Z=3$), ^{9}Be($N=5$) and ^{11}B($Z=5$) fall into this class. Their spins and magnetic moments indicate $p_{3/2}$ in each case. Experiments on the scattering of protons and neutrons by ^{4}He indicate that the lowest state in the unstable nuclei ^{5}He and ^{5}Li, with 3 odd nucleons, is also $p_{3/2}$.

$6<Z, N<8$

^{13}C($N=7$) and ^{15}N($Z=7$) both have $I=\frac{1}{2}$ and magnetic moments indicating a $p_{1/2}$ state.

$8<Z, N<20$

In this range one expects the $1d_{5/2}$, $2s_{1/2}$ and $1d_{3/2}$ levels. ^{17}O, with 9 neutrons, has $I=5/2$ and its magnetic moment μ lies almost exactly on the Schmidt limit for $d_{5/2}$, while ^{19}F, with 9 protons, has $I=\frac{1}{2}$ and μ almost on the Schmidt limit for $s_{1/2}$. ^{21}Ne ($N=11$) and ^{23}Na ($Z=11$) both have $I=3/2$ and μ indicating $p_{3/2}$ rather than $d_{3/2}$. This is the first anomaly encountered in the scheme. It is interpreted as a breakdown of the single-particle model rather than of the order of levels in the potential well. This, and other similar anomalies, will be discussed in §5.3. Although the one-particle model would have assigned $p_{3/2}$, which has odd parity, to the 11th particle on the basis of I and μ, it is entered in fig. 3 as the open circle at $3/2+$, according to the interpretation $(1s_{1/2})^2\,(1p_{3/2})^4\,(1p_{1/2})^2\,(1d_{5/2})^3$.

^{25}Mg ($N=13$) and ^{27}Al ($Z=13$) are both $d_{5/2}$. ^{29}Si ($N=15$) and ^{31}P ($Z=15$) are both $s_{1/2}$. ^{33}S ($N=17$), ^{35}Cl and ^{37}Cl ($Z=17$) are all $d_{3/2}$. ^{39}K ($Z=19$) is $d_{3/2}$. The spin of the radioactive nucleus ^{35}S ($N=19$) has been measured to be $I=3/2$, consistent with $d_{3/2}$, but its magnetic moment is unknown.

With the exception of the anomaly at 11 this behaviour is consistent with the one-particle model and the proposed ordering of levels. It is interesting that at 9 nucleons both $s_{1/2}$ and $d_{5/2}$ are found, indicating that these one-particle levels are close together in the probably very rounded small well appropriate for ^{17}O and ^{19}F. The first excited state of ^{17}O is $\frac{1}{2}+$ (Ajzenberg and Lauritsen 1952), at the relatively low energy (for such a light nucleus) of 0·875 MeV. As the

potential well increases in radius the $2s_{1/2}$ level appears to move upward relative to $1d_{5/2}$, as indicated by its appearance at 15 after the closure of a $d_{5/2}$ sub-shell at 14, which is one of the magic numbers of the strong spin–orbit interaction B-sequence. From 16 to 20 the $d_{3/2}$ sub-shell is filling.

$20<Z,N<28$

All measured odd nuclei in this range are $f_{7/2}$, with the exception of ^{55}Mn ($Z=25$), whose spin $I=5/2$ and magnetic moment $\mu=3{\cdot}47$ n.m. would be interpreted on the single-particle model as $d_{5/2}$. Its interpretation as $(f_{7/2})^5$ will be discussed later.

$28<Z,N<40$

In this region the Jensen–Mayer scheme predicts $2p_{3/2}$, $1f_{5/2}$ and $2p_{1/2}$. No exceptions to this prediction are found. In fact $p_{3/2}$ ground states are found at 29, 31, 33, 35 and 37, $f_{5/2}$ at 37 and $p_{1/2}$ at 39. Since the $p_{3/2}$ level can at most hold four particles its occurrence as a ground state from 29 to 37 inclusive calls for some explanation. That the relative positions of the $p_{3/2}$ and the $f_{5/2}$ levels change over with increasing nuclear size, though it may be a contributory factor, will not explain it, for whichever level of the two is lower the spin of a nucleus with 33 odd particles would be $f_{5/2}$. We have to assume that the $p_{3/2}$ level is the lower, and to appeal to the higher pairing energy in the $f_{5/2}$ level. Thus though the first nucleon after closing the shell at 28 goes into the $p_{3/2}$ level, when there are two additional nucleons they pair off in $f_{5/2}$ states. For this to be able to happen the $f_{5/2}$ pairing energy must be greater than the $p_{3/2}$ energy by at least twice the difference in energy between the one-nucleon levels. Then when a third nucleon is added (nucleon number 31) it again goes into the lower $p_{3/2}$ level, the configuration being $(f_{5/2})^2\,p_{3/2}$. This process will go on until both the $f_{5/2}$ and the $p_{3/2}$ sub-shells are full (at 38), and then at 39 the state will be $p_{1/2}$, as observed.

As has already been mentioned the one-particle model, supplemented by the concept of pairing energy, is only a very rough empirical rule, and should not be pushed too far. The fact that ^{85}Rb and ^{87}Rb, both with 37 protons, have for their ground states respectively $f_{5/2}$ and $p_{3/2}$ is an indication that the situation is more complex.

$40<Z,N<50$

This is the region of the $1g_{9/2}$ sub-shell. The spins actually found are $g_{9/2}$, $p_{1/2}$ and an 'anomalous' $7/2+$ (i.e. $I=7/2$, even parity). As emphasized by Coryell (1953), no noticeable jump in nuclear stability occurs at 40, from which it follows that the $1g_{9/2}$ level is not much above $2p_{1/2}$. The occurrence of $p_{1/2}$ ground states can therefore be explained by the high pairing energy in the $1g_{9/2}$ level, in the same way as the occurrence of $2p_{3/2}$ in competition with $1f_{5/2}$ previously. The $7/2+$ states do not occur in the ground states of any stable nuclei, but have been established from the study of isomerism in ^{77}Ge, ^{79}Se ($N=45$), ^{79}Kr ($N=43$), ^{81}Kr, and ^{105}Rh ($Z=45$), and in the case of ^{79}Se $I=7/2$ has been confirmed by microwave measurements (Hardy, Silvey and Townes 1952). Like the previous anomalies, they are interpreted as a breakdown of the one-particle model for the configurations $(1g_{9/2})^3$ and $(1g_{9/2})^5$. In all these

nuclei ($40<Z,N<50$) the $1g_{9/2}$, $2p_{1/2}$ and $7/2+$ levels lie in a close group (Goldhaber and Sunyar 1951, Goldhaber and Hill 1952), except that for 41 and 49 nucleons the $7/2+$ state is not found. This absence supports the interpretation as $(1g_{9/2})^n$, since for $n=1$ and 9 (one particle, and one hole) the only level in the configuration is $9/2+$.

$50<Z,N<82$

In this shell the expected states are $1g_{7/2}$, $2d_{5/2}$, $2d_{3/2}$, $3s_{1/2}$ and $1h_{11/2}$. All of these are found except $h_{11/2}$, and no others. Protons and neutrons show differences in order of filling in this shell. The odd proton nuclei ($A=121$–205) are larger than the odd neutron nuclei ($A=91$–137). This, as well as the coulomb effect, favours high spins for protons. Thus, at the beginning of the proton shell, $1g_{7/2}$ and $2d_{5/2}$ are in close competition, while $2d_{3/2}$ does not appear until after both these levels are filled at 64. After that, and possibly even before, it appears that pairwise filling of the $1h_{11/2}$ level is taking place, so that the $1h_{11/2}$ sub-shell becomes filled without ever appearing as a ground state. A tendency in this direction was already noted with $1g_{9/2}$ in the previous shell. In the neutron shell, $2d_{5/2}$ fills first, then $3s_{1/2}$ and $2d_{3/2}$. Neither $1g_{7/2}$ nor $1h_{11/2}$ appears as a ground state, again presumably because of pairing effects. The evidence from the many isomeric levels known in this region suggests that $1h_{11/2}$ comes above $1g_{7/2}$ and $2d_{5/2}$, but below $2d_{3/2}$ and $3s_{1/2}$.

A notable feature of the neutron shell is the prevalence of $s_{1/2}$ ground states. In view of the many close configurations possible on the individual level model it is very probable that these states are far from being pure $s_{1/2}$ states, and should more reasonably be denoted just $1/2+$. The same remark no doubt applies with equal force to most of the other ground states.

$82<Z,N<126$

Data are not very plentiful in this shell. The expected states are $1h_{9/2}$, $2f_{7/2}$, $2f_{5/2}$, $3p_{3/2}$, $3p_{1/2}$ and $1i_{13/2}$. Because of the pairing effect one would not expect any $i_{13/2}$ ground states, nor are any known. Examples of all the others are found (fig. 3). The points for ^{227}Ac ($Z=89$, $I=3/2$), ^{231}Pa ($Z=91$, $I=3/2$), ^{237}Np ($Z=93$, $I=5/2$) and ^{241}Am ($Z=95$, $I=5/2$) have not been shown in fig. 3, as no measurement of their magnetic moment has yet been made, nor is there any other indication of the parity of these states.

$N>126$

Information beyond 126 is almost totally lacking. The only known spin is $I=5/2$ for ^{235}U ($N=141$). It is believed on the basis of its β-decay that the spin of ^{209}Pb ($N=127$) is either 11/2 or 9/2, corresponding to the expected $1i_{11/2}$ or $2g_{9/2}$, but at present the evidence is quite inconclusive. The existence of an isomer in ^{211}Po ($N=127$) suggests that a competition between $3p_{1/2}$ and either $2g_{9/2}$ or $1i_{11/2}$ is taking place, similar to the competition between $2p_{1/2}$ and $1g_{9/2}$ in the 40–50 shell, but this is still very uncertain. A detailed study of the order in which the sub-shells fill has been made by Klinkenberg (1952), from whom fig. 7, showing the order of filling separately for protons and neutrons, has been taken.

§ 5. Calculation of Energy Levels

5.1. *jj*-Coupling

The Jensen–Mayer scheme which we have been discussing postulates strong spin–orbit interaction, the one-nucleon levels of a given l being split into two widely separated j-levels. The application of the individual level model in these circumstances classifies the configurations in terms of the one-nucleon j-levels. This classification is known as jj-coupling.

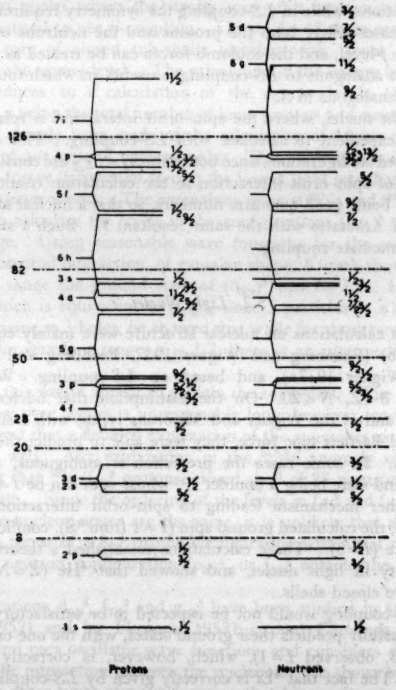


Fig. 7. Proton and neutron levels (*after* Klinkenberg 1952). (Notation of levels different from text.)

(*Reproduced, by permission, from* '*Reviews of Modern Physics*'.)

The position is to be contrasted with the individual level model applied to atoms (more particularly light atoms), in which the splitting of a one-electron doublet $j = l \pm \frac{1}{2}$ is small in comparison with the effects arising from the coulomb repulsion between the electrons in a configuration. In atomic structure one therefore classifies the one-electron levels according to l and treats spin–orbit interaction as a small perturbation. The exclusion principle correlates the total

spin S with the spatial symmetry character of the wave functions, which in turn classifies the types of levels (terms). Total spin is then a constant of the motion, or 'good quantum number'. The total orbital angular momentum L is also a good quantum number, so that a state is characterized by L and S. This state of affairs is known as Russell–Saunders or LS-coupling.

In some respects the mathematical theory of jj-coupling is simpler than of LS-coupling, provided one is entitled to treat protons and neutrons separately, since the exclusion principle requires total antisymmetry in terms of the j sub-level wave functions, while in LS-coupling the symmetry requirements are more complex. In cases where both the protons and the neutrons outside the core are in the same j-level, and the coulomb forces can be treated as a perturbation, a generalization analogous to LS-coupling is useful, in which total isobaric spin T plays a role analogous to S.

In very light nuclei, where the spin–orbit interaction is relatively weak, it may be more expedient to calculate with LS-coupling. This corresponds to mixing jj-coupled configurations since both values $j = l \pm \frac{1}{2}$ are considered together. The inclusion of spin–orbit interaction in the calculation results in neither L or S separately being good quantum numbers, so that a nuclear state is made up of a number of LS-states with the same resultant I. Such a state of affairs is known as intermediate coupling.

5.2. *Light Nuclei*

The earliest calculations on nuclear structure were mainly confined to light nuclei ($Z, N \leqslant 8$) (Feenberg and Wigner 1937, Feenberg and Phillips 1937, Hund 1937, Wigner 1937a), and based on LS-coupling. Wigner (1937b) also considers $8 < Z$, $N \leqslant 20$. On the assumption that nuclear forces were mainly central and of the Wigner and Majorana types, with a small amount of Heisenberg and Bartlett type force, they were able to predict the ground levels of these nuclei. In some cases the prediction is ambiguous, as in ^{7}Li, the calculated ground state being a doublet ^{2}P, whose spin can be $I = 1/2$ or $I = 3/2$, and some further mechanism leading to spin–orbit interaction is necessary. In one case, ^{10}B, the calculated ground spin ($I = 1$ from ^{3}S), conflicts with experimental evidence ($I = 3$). These calculations established a theoretical basis for the α-periodicity in light nuclei, and showed that ^{4}He ($Z = N = 2$) and ^{16}O ($Z = N = 8$) were closed shells.

Although jj-coupling would not be expected to be satisfactory in such light nuclei, it successfully predicts their ground states, with the one exception of ^{6}Li (predicted $I = 3$, observed $I = 1$), which, however, is correctly predicted on LS-coupling. The fact that ^{6}Li is correctly given by LS-coupling and ^{10}B by jj-coupling points to the necessity of basing the calculations on intermediate coupling. Inglis (1953) gives a very full account of the theory of light nuclei.

The extension of LS-coupling calculations beyond ^{16}O requires a classification of the states of configurations d^n or f^n according to symmetry types. Such work has been done by Racah (1949, where earlier references will be found) for atoms, and extended by Jahn (1950, 1951), Jahn and van Wieringen (1951), Swiatecki (1951), Flowers (1952a). No very detailed application of this work has been made.

One feature common to all these calculations may be mentioned. It is that the ground state of all even–even nuclei turns out to be $^1S_0+$ (i.e. $I = 0$, even

parity). This is in agreement with observation. This feature is related to the fact that two nucleons in the same orbital (l) have the greatest probability of being close together when their angular momenta couple to zero resultant. Since the forces between them are attractive this results in a lowered energy. If the two nucleons are identical, the exclusion principle automatically assures that their spins pair off also in such a state.

5.3. *The Configuration j^n*

If we turn to nuclei where the strong spin–orbit interaction may not be neglected, it is natural to use jj-coupling. Up to $A=48$ there are many nuclei composed of a core of closed sub-shells together with 'loose' protons and neutrons both in the same j-level. The calculation of the energy levels of such a configuration reduces to a calculation of the energy levels of the simple configuration j^n, n being the total number of loose protons and neutrons.

Mention has already been made of the calculation of Mayer (1949) for the configuration j^n of one type of nucleon only. This showed that in the limit of very short range forces (δ-function forces) the lowest level was that with $I=0$ or j according as n is even or odd. The anomalous spins of ^{23}Na and ^{55}Mn prompted Kurath (1950) to calculate the levels in the configurations $(d_{5/2})^3$ and $(f_{7/2})^5$ as a function of range. Using reasonable wave functions for the nucleons and a simple central potential interaction, of gaussian shape, Kurath showed that for a sufficiently long range the ground state of $(d_{5/2})^3$ was $I=3/2$. For $(f_{7/2})^5$, the calculation of which is equivalent to $(f_{7/2})^3$, since 5 particles in a sub-shell of 8 sub-levels is the same as 3 holes, he showed that while for short ranges the lowest level is 7/2, and for very long ranges it is 3/2, there is an intermediate range for which it is 5/2.

At first sight this would appear to explain the anomalies. The range of force at which the cross-over occurs is unreasonably long, however, and more recent calculations suggest that a detailed explanation of the anomalies must depend on other factors as well. Any explanation of the ^{55}Mn anomaly must take into account that fact that ^{51}V ($f_{7/2}^3$ protons, closed shell of 28 neutrons) has $I=7/2$ in the ground state. Since the ordering of the levels in $f_{7/2}^3$ and $f_{7/2}^5$ is the same, provided the one-nucleon states are the same, it is unlikely that an explanation treating ^{55}Mn in terms of a $f_{7/2}^5$ proton configuration without at the same time considering the neutron configuration ($p_{3/2}^2$, or $f_{5/2}^2$ outside the core) will be successful.

The configurations $d_{5/2}^3$, $f_{7/2}^3$ and $g_{9/2}^3$ have been studied in some detail by Talmi (1951, 1952) and by Flowers (1952b, c) and Edmonds and Flowers (1952a, b). Talmi uses oscillator wave functions and considers different types of potential for the interaction between the nucleons. He shows by a numerical comparison with other types of wave function that his results are insensitive to the choice of wave function. For short ranges of force, likewise, the results are rather insensitive to the form of potential, but at the longer ranges there are qualitative differences between potentials with a singularity at the origin (such as a coulomb or Yukawa potential) and smooth potentials (e.g. gaussian). For $f_{7/2}^3$, for example, in a singular potential the lowest level is never 5/2, while in the gaussian potential, as already shown by Kurath, it is 5/2 at an intermediate range. Talmi concluded that for no reasonable central potential does $I=3/2$ come lowest in $d_{5/2}^3$, nor $I=5/2$ in $f_{7/2}^3$, although they come close to the ground

level. Some additional, possibly quite weak, effect must therefore enter into the explanation.

For $g_{9/2}^3$, however, Talmi (1952) calculates that $I=7/2$ can reasonably come below $I=9/2$, provided a singular potential is used. Apparently the most important factor in the competition between 9/2 and 7/2 is the relative amount of Majorana (exchange) type and Wigner (ordinary) type force. For pure Wigner force the order of the levels (at a realistic range, potential $\propto(r/r_0)\exp(-r^2/r_0^2)$) is 7/2, 9/2, 5/2, ..., the spacing between the latter two being about four times that between the former, while for pure Majorana forces the order is 9/2, 3/2, ... For equal mixtures the order is 7/2, 9/2, ..., their spacing now being relatively greater than for Wigner forces. There is therefore a reasonable expectation that a closer study of the levels of the $g_{9/2}^3$ configuration will throw some light on the nature of nuclear interactions.

The general problem of the classification of states in an arbitrary configuration j^n of protons and neutrons has been treated by Flowers (1952b), extending the group-theoretical methods of Racah (1949), Jahn (1950, 1951) and Swiatecki (1951). Flowers introduces two new quantum numbers, the seniority s and the reduced isobaric spin t, in addition to the total angular momentum I and isobaric spin T. These are not constants of the motion in any actual nucleus, but for sufficiently short range interactions there are sufficiently 'good' quantum numbers to furnish an extremely useful classification (Edmonds and Flowers 1952a). The lowest level of a configuration has the lowest permitted seniority s ($s=0$ if n is even, $s=1$ if n is odd), and has spin $I=0$ or j, correspondingly, as already established by Mayer (1950b) for a δ-function potential. The next levels form a group, with seniority 2 or 3, according to the parity of n, with a variety of possible values of I. The gap between the ground level and the group of next higher seniority can be interpreted as the pairing energy already defined in §4.

Flowers gives the detailed classification of $(3/2)^n$, $(5/2)^n$, $(7/2)^n$ (1952b), and of $(9/2)^n$ with particles of one kind only (1952c). Edmonds and Flowers (1952b) have computed the energy levels of a number of the possible configurations $p_{3/2}^n$, $d_{5/2}^n$, $f_{7/2}^n$ and (Flowers 1952c) $g_{9/2}^n$, using oscillator wave functions and a gaussian potential of the Rosenfeld exchange type (Rosenfeld 1948). They have also computed the energy levels of $d_{5/2}^n$ with a Yukawa potential $r^{-1}\exp(-r/r_0)$. Their conclusions agree with those of Talmi in the cases treated by Talmi, except for $g_{9/2}^n$, for which Flowers never finds $I=7/2$ lowest. The discrepancy is to be ascribed to the non-singular, gaussian potential used by Flowers, Talmi having stressed how important the $1/r$ singularity is in the ordering of the levels.

§6. The Energy Levels of Even–Even Nuclei

It has already been mentioned that without known exception the ground state of an even–even nucleus is $I=0$. All calculations along the lines already mentioned lead to this conclusion (see also Pryce 1952). The calculations of Edmonds and Flowers (1952a, b) show that (for sufficiently short-range interactions) the next highest states of a configuration j^n (n even, made up of an even number of loose protons and an even number of loose neutrons) are in the order

$$I=2,4,\ldots 2j-1,$$

after which comes an irregular group of states of higher seniority and/or higher isobaric spin.

Experimentally, the first excited state, with only two known exceptions out of 68 cases, is $I=2(+)$ (Scharff-Goldhaber 1953). The two exceptions are ^{16}O ($Z=N=8$) and ^{32}Ge ($Z=32$, $N=40$), with $I=0(+)$. Both these nuclei have closed shells or sub-shells of both protons and neutrons. ^{32}Ge, with 32 protons, has a closed sub-shell of 4 $2p_{3/2}$ protons in addition to the closed shell at 28. The explanation of these two exceptional cases on the basis of the shell model is straightforward. As the ground configuration is a closed core, it contains only one level, $I=0(+)$. The next level must belong to an excited configuration. At first sight it would appear that this was to be formed by taking one nucleon from a closed shell and lifting it to a state in the next shell. It is possible, however, that because of high pairing energy, a lower level results by taking out two nucleons from a closed shell, leaving it with resultant angular momentum 0, and pairing them off with zero resultant in the next shell. In this way a $0+$ level results.

According to the calculations of Edmonds and Flowers the excitation energy of the first $(2+)$ level in a pure j^n configuration of protons and neutrons should not vary very markedly with n, provided A is not too small, depending mainly on a slow change in the nuclear radius. This prediction is not very easy to test, however, because protons and neutrons occupy different levels beyond $A=48$, and below this the pure configurations do not occur in many even–even nuclei. The best test is the $1f_{7/2}$ shell, with the nuclei ^{42}Ca, ^{44}Ca, ^{46}Ti, ^{48}Ti, ^{50}Ti, ^{52}Cr. Experimentally, the excitation energies of the first level of these nuclei lie in the range 1·1–1·6 MeV (Horie, Umezawa, Yamaguchi and Yoshida 1951, Stähelin and Preiswerk 1951, Scharff-Goldhaber 1953).

A plot of the energy of the first excited level against nucleon number shows a striking correlation with the closed shells. There is a general decreasing trend with increasing mass number, which a simple model shows can be represented by the dependence $A^{-2/3}$ (Stähelin and Preiswerk 1951). If this factor is taken out, the reduced excitation energy shows sharp peaks at the magic numbers (of either Z or N) with a shallow valley in between. Doubly magic nuclei have abnormally high peaks. Stähelin and Preiswerk (1951) have shown the remarkable fact that, except for the doubly magic nuclei, the reduced excitation energy can be very satisfactorily represented as the sum of two functions $f(Z)$ and $f(N)$ of the proton and neutron number respectively. This function (it has essentially the same dependence on Z and N) has peaks at the magic numbers, separated by shallow troughs. No theoretical explanation of this rule is yet known.

The second excited states of even–even nuclei are less well known. They seem to be about 35% $4+$, 30% $2+$ and the remainder miscellaneous.

According to Edmonds and Flowers (1952a, b), in a pure j^n configuration, the $4+$ second excited level should lie rather close to the $2+$ first level. This result is only applicable to pure configurations, and there is strong reason to believe that configurational mixing plays an important part in excited states, so it is not surprising to find that this rule is not well fulfilled. Indeed, Bohr and Mottelson (1953) have pointed out that there are many cases where the level spacing conforms more to the law $E \propto I(I+1)$ with $I=0, 2, 4, \ldots$. These remarkably simple sequences all happen in fairly heavy nuclei well away from closed shells. They are explained by Bohr and Mottelson on the basis of the collective motion (deformable core) model of the nucleus.

§7. Odd–Odd Nuclei

The behaviour of odd–odd nuclei, with one unpaired proton and one unpaired neutron, is somewhat irregular. In the ground state high spins are often encountered. On the whole the ground states can be explained in terms of the one-particle levels assigned to the odd proton and neutron according to the scheme given in §4. If j and j' are the angular momenta of the proton and neutron respectively, the resultant I can take any value between $|j-j'|$ and $j+j'$, always assuming that the odd nucleons combine among themselves to give the usual j and j' (i.e. that the anomalous behaviour found in ^{23}Na etc. is not taking place). To determine which is the ground level involves, in principle, a calculation of the levels of the mixed configuration $j^n j'^{n'}$, n and n' being the numbers of protons and neutrons outside the core. Such calculations have been performed in relatively few cases.

Nordheim (1950) has proposed an empirical rule that j and j' set themselves either parallel or antiparallel, whichever gives the most parallel arrangement of the spins of the proton and neutron. Thus if $j=l+\frac{1}{2}$, $j'=l'+\frac{1}{2}$, or $j=l-\frac{1}{2}$ and $j'=l'-\frac{1}{2}$, then $I=j+j'$. If $j=l+\frac{1}{2}$ and $j'=l'-\frac{1}{2}$, or $j=l-\frac{1}{2}$ and $j'=l'+\frac{1}{2}$, then $I=|j-j'|$. A reason for this rule can be given on the grounds that the strong Majorana forces favour a parallel alignment of spins, while the probability of the two nucleons being close together has a maximum both for a resultant $L=|l-l'|$ and for $L=l+l'$, although the former maximum is usually much stronger. Calculations made by Pryce (1952) for the configuration jj' on the basis of a δ-function interaction do not wholly support this, though they do show that the ground state is either $I=|j-j'|$ or $I=j+j'$ (though with $j=j'=3/2$, $I=1$ is equal with $I=3$). The same conclusions hold for $j^{-1}j'^{-1}$. For $j^{-1}j'$ or jj'^{-1} however the ground state is $I=j+j'-1$ (unless j or $j'=\frac{1}{2}$, in which case the Nordheim rule is always correct). This was shown by Kurath (1952).* These two extreme cases show that no such simple rule as Nordheim's can have universal validity.

King and Peaslee (1953) give empirical evidence to show that when protons and neutrons are filling the same shell the spins of the odd–odd nuclei increase regularly from 1 to $2j$. For instance in the $1p_{3/2}$ shell, ^{6}Li has $I=1$, ^{8}Li and ^{8}B probably $I=2$ and ^{10}B, $I=3$. Again, for somewhat heavier nuclei, in the $1f_{7/2}$ shell, ^{44}Sc (2 or 3), ^{46}Sc (4), ^{48}V (4 or 5), ^{50}V (6), ^{52}Mn (6 or 7). Such a behaviour is in sharp contrast with Nordheim's rule except at the end of a sequence.

In a configuration of two like particles the exclusion principle forbids half of the states (in particular the parallel arrangement $I=2j$ when both are in the same j-level), but this does not operate when the particles are of different kinds. Odd–odd nuclei consequently tend to have a more complex arrangement of low-lying energy levels than even–even nuclei. The possible competition of states of high spin with states of low spin leads to the frequent occurrence of isomerism (§8) in odd–odd nuclei. An example is ^{210}Bi, one state of which (RaE) has low spin ($I\leqslant 2$) and decays predominantly by β-emission, while the other state decays predominantly by α-emission and probably has $I\geqslant 8$ (Pryce 1952).

* This result is at variance with the original result stated by Pryce (1952) but is in accordance with the corrigendum.

§ 8. Isomerism

The close relation between shell structure and the incidence of isomerism has often been noted (Feenberg and Hammack 1949, Nordheim 1949). If, purely as a matter of convenience, we define an isomeric transition to mean a radiative transition in which there is a change in I of at least three units (this is quite arbitrary) then isomerism first occurs systematically in connection with the $1g_{9/2}$ shell. The close proximity of the $2p_{1/2}$ and $1g_{9/2}$ levels, of different parities and differing in angular momentum by four units, leads to the frequent occurrence of isomers of the M4 type (magnetic 2^4-pole). Transitions between $7/2+$ levels arising from $1g_{9/2}^n$ configurations and $1/2-$ levels from configurations involving $2p_{1/2}$ similarly give rise to a group of E3 isomers. The work of Goldhaber and Sunyar (1951) in classifying isomeric transitions has played a considerable part in unravelling the spins and parities of many nuclear levels, and, in particular, has established the occurrence of low $7/2+$ levels in the $1g_{9/2}^n$ shell.

The proximity of $1h_{11/2}$ and $2d_{3/2}$ or $2d_{5/2}$ in the next shell (50–82) and of $1i_{13/2}$ and $2f_{5/2}$ or $2f_{7/2}$ in the shell beyond that (82–126) leads to two further groups of M4 and E3 isomers, mainly in odd nuclei.

Other types of isomers are found scattered more irregularly in the nuclear periodic table, mainly among odd–odd nuclei. Of these one may quote ^{58}Co ($Z=27$, $N=31$). This may be regarded as one proton hole in the $1f_{7/2}$ sub-shell and one neutron hole in the $2p_{3/2}$ sub-shell. Using the results established by Pryce (1952) for very short range forces in a configuration of one proton and one neutron (or equivalently holes) one finds that the lowest levels are $I=2$ and $I=5$, these being so close together that without more accurate knowledge of the interactions (in this case the amount of Heisenberg and Bartlett force relative to Wigner and Majorana force) it is impossible to decide which is the lower. Experimentally two levels separated by 25 kev, to which the spins $2+$ (ground level) and $5+$ are assigned on the basis of isomeric behaviour and beta-decay, are found. The assigned configuration may be wrong, however, another possibility being $f_{7/2}^{-1} f_{5/2}^3$, which may compete strongly because of the similarity of the $f_{7/2}$ and $f_{5/2}$ (radial) wave functions. Yet another possibility is $(f_{7/2}^{-2} f_{5/2})(f_{5/2}^3)$, the first bracket referring to protons, the second to neutrons. Goldhaber (1953) and de Shalit and Goldhaber (1953) have pointed out the stabilizing effect obtained by transferring one type of nucleon to another level where it can interact strongly with nucleons of the other sort. The measured magnetic moment of the ground state of ^{58}Co (Daniels *et al.* 1952) would be in better agreement with the latter configurations.

Goldhaber (1953) has used the idea just mentioned to make an interesting suggestion concerning the isomerism of ^{93}Mo ($Z=42$, $N=51$). He suggests that a $1g_{9/2}$ proton is lifted to $1g_{7/2}$ to couple with the odd neutron (assumed to be in a $1g_{7/2}$ state). He proposes to regard the three stage γ-ray cascade (E4 – E2 – E2) as transitions between states in which an even–even core (42 protons + 50 neutrons) jumps from $8+ \to 4+ \to 2+ \to 0+$, the neutron remaining in a $g_{7/2}$ state the whole time. This process he calls 'core isomerism'. It should be remarked in passing that Goldhaber's assignment of $1g_{7/2}$ to the 51st neutron is in conflict with the evidence from stable nuclei, in which the 51st neutron is always in a $2d_{5/2}$ state, but this objection does not invalidate his suggestion.

Very few even–even nuclei show isomerism. This is in line with the

considerations of § 6. The only ones known at present are ^{180}Hf ($Z=72$, $N=108$), ^{202}Pb ?, ^{204}Pb and ^{206}Pb ($Z=82$, $N=120, 122, 124$). The decay schemes of ^{180}Hf (Goldhaber and Hill 1952) and ^{202}Pb (Hopkins 1952) are not known, and the isotopic assignment of the latter is still uncertain. In ^{204}Pb the transition is of E5 type from $7-$ to $2+$ (Goldhaber and Hill 1952). In ^{206}Pb it is E3, from $7-$ to $4+$ (Alburger and Pryce 1953). The $7-$ level in ^{206}Pb almost certainly arises from a configuration $1i_{13/2}^{-1}\,3p_{1/2}^{-1}$, and in ^{204}Pb from configurations involving $1i_{13/2}$. The anomalous isomerism of the lead isotopes can be connected with their proximity to the double closed shell ^{208}Pb. The occurrence of isomerism in ^{180}Hf has not been explained.

The reader is referred to the review by Goldhaber and Hill (1952) for a fuller account of isomerism in relation to shell structure.

§ 9. Magnetic Moments and Quadrupole Moments

9.1. *Magnetic Moments of Odd Nuclei*

As a general rule the magnetic moments of odd nuclei fall considerably off the Schmidt lines. Several types of explanation have been proposed for these deviations.

According to meson theories of nuclear interactions the magnetic moment of a nucleon is partly intrinsic and partly due to a 'meson cloud' surrounding the nucleon. The exchange of mesons which gives the exchange forces also leads to an exchange current and magnetic moment. The theory of the effect is still uncertain, but it is believed to be small, and insufficient to account for the deviations from the Schmidt limits (Sachs 1948, Osborn and Foldy 1950, Spruch 1950) except in ^{3}H and ^{3}He. Ross (1952), however, claims larger deviations.

It has also been suggested (Bloch 1951, Miyazawa 1951, de Shalit 1951, 1953) that the anomalous magnetic moment of a nucleon becomes quenched when it interacts with others in a nucleus. This would explain the direction of the deviations. If this is the correct explanation, however, then heavy nuclei such as ^{197}Au whose moments lie close to the Schmidt limits are difficult to explain.

On the other hand the magnetic moment of a nucleus would only be expected to lie on the Schmidt limits if the one-particle model were strictly valid. On the more general, but still approximate, individual level model, the magnetic moment is in most cases expected to deviate from the Schmidt value. Only when the even type of nucleons fill a shell (or sub-shell) does this model predict moments in agreement with the one-particle model. Such nuclei, provided they are sufficiently light, do in fact show smaller deviations than others. For example ^{3}H ($N=2$), ^{3}He ($Z=2$), ^{13}C ($Z=6$), ^{15}N ($N=8$), ^{17}O ($Z=8$) lie rather close to the Schmidt lines. The somewhat heavier nuclei ^{37}Cl ($N=20$), ^{39}K ($N=20$) and ^{51}V ($N=28$), however, already show considerable deviation.

In light nuclei where isobaric spin is a good quantum number (coulomb forces not too important) a straightforward calculation of the magnetic moment is possible, on the basis of the known symmetry of the wave function (Umezawa 1952, Mizushima and Umezawa 1951, 1952, Flowers 1952 d). A simple example will make the principle clear. Consider ^{7}Li in jj-coupling. It has one proton and two neutrons outside the closed ^{4}He core, occupying the configuration $(1p_{3/2})^3$. The wave function of the ground has definite symmetry properties which specify it completely. It can be decomposed into a part representing the proton coupled to

the two neutrons in a state of zero resultant angular momentum, and another part in which the proton angular momentum is vector coupled to a resultant neutron angular momentum of 2. The weights of these two states are respectively 5/6 and 1/6. The magnetic moments associated with each can be calculated by the usual Landé-type formula and are respectively $\mu_p+1=3{\cdot}79$ and $(\mu_p+4\mu_n+1)/5=-0{\cdot}77$ n.m. The expected moment of the ^{7}Li state is thus $\mu=(5/6)\ 3{\cdot}79-(1/6)\ 0{\cdot}77=3{\cdot}03$ n.m. This calculation has been given merely to illustrate the principle in a particularly simple example. It should not be taken as a serious estimate of the moment of ^{7}Li, which is too light for *jj*-coupling to be valid. The observed value μ is 3·26 n.m. The calculation is instructive in showing that, although five sixths of the wave function is as postulated by the one-particle model, the remaining one sixth can still be responsible for a considerable change in the magnetic moment.

This type of calculation can be applied to all nuclei in which protons and neutrons are filling the same level. The results are in considerably better agreement with observed moments than those of the single-particle model.

When protons and neutrons occupy different incomplete sub-shells a knowledge of symmetry is not enough to determine the wave function type uniquely. It must nevertheless be expected that the theoretical moments differ from the Schmidt limits. Mixing of configurations also modifies the magnetic moment. Deviations from the Schmidt limits must therefore be expected in all nuclei.

The deviations can be described in terms of a 'polarization', or a tidal deformation, of the core. In the example just discussed, the state in which the two neutrons have resultant 2 can be considered as a deformed neutron core. The deformed core then possesses an angular momentum which precesses about the resultant I of the nucleus, and gives a contribution to the magnetic moment. This type of explanation has been discussed particularly by Rainwater (1950), A. Bohr (1951), Foldy and Milford (1950) and Feenberg and Hammack (1951). Core deformation does not account for a deviation of magnetic moments of nuclei with $I=\frac{1}{2}$, because the lowest angular momentum of the deformed core is 2, which cannot be coupled with $\frac{1}{2}$ to give $I=\frac{1}{2}$. An explanation of the deviations of such nuclei in terms of other types of core excitation has been discussed by Blin-Stoyle (1953).

9.2. *Magnetic Moments of Odd–Odd Nuclei*

The moments calculated on the basis of observed spin I and the one-particle levels assigned to the proton and neutron separately are generally in better agreement with observation than for odd–even or even–odd nuclei. This point has been discussed by Feenberg (1949), de Shalit (1953), Schwartz (1953), on the basis of a quenching of intrinsic moment, who point out that the quenching effects would be expected to compensate in odd–odd nuclei. A detailed discussion on the basis of the individual level model is so far lacking, but it seems probable that the effects held responsible for the deviations in odd-A nuclei would also compensate in odd–odd nuclei.

Quadrupole moments.

It has been pointed out by Horie and Yoshida (1951) and Flowers (1952 d) that the quadrupole moments expected in light nuclei, on the basis of the charge independent forces, are of the same sign and same order of magnitude for odd-neutron as for odd-proton nuclei, a result at variance with the predictions of the

single-particle model. The qualitative features observed, namely negative quadrupole moments at the beginning of a shell, positive at the end, fit in with the calculations. In the middle of a shell the quadrupole moment can be considerably larger than that predicted on a single-particle model. Even so, the very large quadrupole moments observed in the rare-earth region cannot be explained in terms of particles in a single configuration. They have been interpreted by Rainwater (1950) as definitive evidence of a non-spherical shape of the nuclear core, and find a natural explanation in collective models of the nucleus (Van Wageningen and de Boer 1952). A calculation by Pfirsch (1952) for a single configuration, introducing a deformability parameter, is intermediate in viewpoint.

§ 10. Limitations of the Individual Level Model—Collective Models

10.1. *Configurational Mixing*

It has already been pointed out that on the individual level model a single configuration will only give a reasonable representation if the matrix elements between two configurations of the interaction are small compared with their energy separation. Otherwise several configurations must be considered at the same time, the resulting eigenvalue problem for the energy levels at once becoming vastly more complex. It is therefore of considerable importance to make an estimate of the inter-configuration matrix elements. With the exception of calculations on very light nuclei (see Inglis (1953) for references) no systematic study has been published. The matrix elements are bound to be of the same general order of magnitude as the diagonal matrix elements occurring in the determination of energy levels in a single configuration, though on the average probably somewhat smaller. The pairing energy of a pair of nucleons in a given level may be taken as an upper limit, this being the diagonal element in $j^2(I=0)$. Estimates of diagonal matrix elements and pairing energies for heavy nuclei have been made by Pryce (1952) and Edmonds (1953), and for light elements by Hitchcock (1953). Pairing energies are in the general region of a few Mev in all nuclei, with a gradual decrease by a factor of about 3 from $A=20$ to $A=200$ according to Hitchcock. The off-diagonal matrix elements may therefore be estimated to be of the order of magnitude of 100 kev. This is of the same order as (though perhaps smaller than) the spacing between different levels in one shell, so that a single configuration can hardly be expected to be a good approximation in the middle of a shell, where the number of configurations is large, particularly for excited states.

Near a closed shell, particularly for low-lying states, the position is relatively favourable, and one might expect to make reasonable qualitative statements about the classification of levels starting from a single conhguration. A study of the possible low levels of ^{210}Bi ($Z=83=82+1$, $N=127=126+1$) has been made by Pryce (1952), from which it appears that levels from different configurations overlap considerably, and that even this apparently simple problem presents many difficulties. A classification of the levels of ^{206}Pb ($Z=82$, $N=124=126-2$), on the other hand, though encountering difficulties from mixed configurations, appears to be successful (Alburger and Pryce 1954).

The situation is probably more favourable in the $1f_{7/2}$ shell, which is fairly well separated in energy from the shells above and below. A doubtful point, however, is the extent to which jj-coupling is a good starting point. Quantitative calculations

in medium nuclei indicate that *jj*-coupling is broken down to an appreciable extent. This at once complicates the problem again.

The classification of the levels observed in isomeric nuclei poses some interesting questions. In many instances several other levels beside the isomeric pair are known, and their spins and parities are reasonably established. It is usual to interpret them as one-particle levels, but clearly even this interpretation is only indirectly connected with the ordering of one-nucleon levels in the effective potential. As an example let us consider an odd N nucleus in which the $1g_{7/2}$ sub-shell is filling, the $2d_{5/2}$ sub-shell being full. The available levels above $1g_{7/2}$ are $1h_{11/2}$, $2d_{3/2}$ and $3s_{1/2}$, which are to be considered empty in the ground state of the nucleus. On the refined version of the one-particle model this is $7/2+$. Excited levels can be formed by taking the unpaired neutron and lifting it to one of the empty levels, thus giving $11/2-$, $3/2+$ and $1/2+$, no pairing energies being involved. Other excited levels can be formed by promoting a $2d_{5/2}$ neutron to $1g_{7/2}$, thereby expending one $2d_{5/2}$ pairing energy and gaining in return one $1g_{7/2}$ pairing energy, or by promoting a $2d_{5/2}$ and a $1g_{7/2}$ neutron to pair off in $1h_{11/2}$. Both these levels are $5/2+$, and if the difference in pairing energy between $1h_{11/2}$ and $1g_{7/2}$ is nearly equal to twice the one-particle energy gap between them, these levels will be close together. Table 2 will illustrate this point. We denote the

Table 2

Occupation of :	$2d_{5/2}$	$1g_{7/2}$	$1h_{11/2}$	$2d_{3/2}$	$3s_{1/2}$	Spin and parity	Energy
	6	3	0	0	0	$7/2+$	0
	6	2	1	0	0	$11/2-$	E_1-E_0
	6	2	0	1	0	$3/2+$	E_2-E_0
	6	2	0	0	1	$1/2+$	E_3-E_0
	5	4	0	0	0	$5/2+$	$E_0-E_{-1}+\pi_{-1}-\pi_0$
	5	2	2	0	0	$5/2+$	$2E_1-E_0-E_{-1}+\pi_{-1}-\pi_1$
	4	5	0	0	0	$7/2+$	$2E_0-2E_1+\pi_{-1}-\pi_0$
	4	3	2	0	0	$7/2+$	$2E_1-2E_{-1}+\pi_{-1}-\pi_1$
	6	3	0	0	0	$5/2+$	?

one-particle energies E_{-1}, E_0, E_1, E_2, E_3 and the corresponding pairing energies π_{-1}, π_0, π_1, π_2, π_3, and for the sake of definiteness we consider a nucleus with three neutrons outside the $2d_{5/2}^6$ core. The energies are referred to the $5/2+$ ground state. Additional levels can be formed by promoting two neutrons from $2d_{5/2}$ to $1g_{7/2}$, or to $1h_{11/2}$. These will be $7/2+$ levels. One has also to consider the possibility, not accounted for on the one particle model, that $1g_{7/2}^3$ $I=5/2$ may be fairly low. Yet other combinations are possible. Rough though this model is, it reveals the complexity to be expected of the classification. Viewed from the more comprehensive standpoint of the individual level model it suggests that the higher levels of the configuration $1g_{7/2}^3$ must mix with these other types of level. The effect of configurational mixing is to push apart levels of the same I and parity. In this way the $5/2+$ and $3/2+$ levels, which are probably fairly low in $1g_{7/2}^3$ get pushed down. The question naturally arises of how justified one is in denoting even the ground state by the symbol $g_{7/2}$.

Wigner (1951) has made a rough estimate of how pure a configuration a closed shell is, the case being the purest one would expect. His estimate is that only 25% of the wave function is represented by the nominal configuration. Such an

estimate makes one wonder how any validity at all can be accorded to the individual level model. The answer might be that although the nominal configuration is only 25%, the admixed configurations are very many in number, each present in a much smaller proportion, so that the nominal configuration dominates the properties of the nucleus. Nevertheless the conclusion is inescapable that, in a filling shell with many near one-particle levels, configurational mixing plays an important part.

10.2. *Collective Models*

From this point of view a new line of attack on the problem is desirable. This is furnished by the collective model of A. Bohr (1951), partly based on the spheroidal model of Rainwater (1950), and discussed also by Hill and Wheeler (1953), and others (e.g. Gallone and Salvetti 1951). In this model, the potential field in which the nucleons move is not considered as spherically symmetrical, but as arising from a deformable 'core'. The most easily excited deformations can be likened to the ellipsoidal deformations of the surface of a liquid drop. The surface deformations can rotate without the nuclear matter rotating as a whole, and are capable of carrying angular momentum. They are strongly coupled with the motions of individual nucleons, through the action of the effective, non-spherical, potential, so that the resultant I of the nucleus is made up of two strongly coupled parts.

Quantitative predictions of relations between quadrupole moments, deviations of magnetic moments from the Schmidt limits, and the low energy levels of even–even nuclei can be derived (Bohr and Mottelson 1953). A further discussion of the collective model is outside the scope of this Report, and the reader is referred to the authors quoted above.

10.3. *The Spin–Orbit Interaction*

One further point which should be discussed concerns the origin of the strong spin–orbit interaction which is involved in the Jensen–Mayer level scheme. Our knowledge of the origin of nuclear forces is still too slight to give a basis for it, but so far no theory of nucleon–nucleon interactions has given any wholly acceptable justification for it (see Gaus 1949, 1952, however). Attempts have been made to account for the effect in terms of tensor forces (Keilson 1951, Höcker 1952). It is best to regard it at the moment as an observed fact whose origin is unexplained.

References

Alburger, D. E., and Pryce, M. H. L., 1953, *Phys. Rev.*, **92**, 514; 1954, *Ibid.*, in the press.

Ajzenberg, F., and Lauritsen, T., 1952, *Rev. Mod. Phys.*, **24**, 321.

Bartlett, J. H., 1932, *Nature, Lond.*, **130**, 165, *Phys. Rev.*, **41**, 370, *Ibid.*, **42**, 145.

Béné, G.-J., 1952, *J. Phys. Radium*, **13**, 161.

Blin-Stoyle, R. J., 1953, *Proc. Phys. Soc.* A, **66**, 1158.

Bloch, F., 1951, *Phys. Rev.*, **83**, 839.

Bohr, A., 1951, *Phys. Rev.*, **81**, 134; 1952, *Dan. Mat. Fys. Medd.*, **26**, no. 14.

Bohr, A., and Mottelson, B. R., 1953, *Phys. Rev.*, **89**, 516, *Ibid.*, **90**, 587.

Bohr, N., 1936, *Nature, Lond.*, **137**, 344.

Bouchez, R., Robert, J., and Tobailem, J., 1953, *J. Phys. Radium*, **14**, 72.

Coryell, C. D., 1953, *Annual Review of Nuclear Science*, **2**, 305.

Daniels, J. M., Grace, M. A., Halban, H., Kurti, N., and Robinson, F. N. H., 1952, *Phil. Mag.*, **43**, 1297.

Duckworth, H. E., 1952, *Nature, Lond.*, **170**, 158.

EDMONDS, A. R., 1953, *Proc. Phys. Soc.* A, **66**, 793.
EDMONDS, A. R., and FLOWERS, B. H., 1952 a, *Proc. Roy. Soc.* A, **214**, 515; 1952 b, *Ibid.*, **215**, 120.
ELSASSER, W., 1933, *J. Phys. Radium*, **4**, 549; 1934, *Ibid.*, **5**, 389, 635.
FEATHER, N., 1953, *Advances in Physics* (*Phil. Mag.* Suppl.), **2**, 141.
FEENBERG, E., 1949, *Phys. Rev.*, **76**, 1275.
FEENBERG, E., and HAMMACK, K. C., 1949, *Phys. Rev.*, **75**, 1877; 1951, *Ibid.*, **81**, 285.
FEENBERG, E., and PHILLIPS, M., 1937, *Phys. Rev.*, **51**, 597.
FEENBERG, E., and WIGNER, E. P., 1937, *Phys. Rev.*, **51**, 95.
FLOWERS, B. H., 1952 a, *Proc. Roy. Soc.* A, **210**, 497; 1952 b, *Ibid.*, **212**, 248; 1952 c, *Ibid.*, **215**, 398; 1952 d, *Phil. Mag.*, **43**, 1330; 1952 e, *Progress in Nuclear Physics*, **2**, 235.
FOLDY, L. L., and MILFORD, F. J., 1950, *Phys. Rev.*, **80**, 751.
GALLONE, S., and SALVETTI, C., 1951, *Nuovo Cim.*, **8**, 960.
GAUS, H., 1949, *Z. Naturforsch.*, **4a**, 721; 1952, *Ibid.*, **7a**, 44.
GLUECKAUF, E., 1948, *Proc. Phys. Soc.*, **61**, 25.
GOLDHABER, M., 1953, *Phys. Rev.*, **89**, 1146.
GOLDHABER, M., and HILL, R. D., 1952, *Rev. Mod. Phys.*, **24**, 179.
GOLDHABER, M., and SUNYAR, A. W., 1951, *Phys. Rev.*, **83**, 906.
GORDY, W., 1949, *Phys. Rev.*, **76**, 139.
GREEN, A. E. S., and EDWARDS, D. F., 1953, *Phys. Rev.*, **91**, 46.
GRIFFITHS, J. H. E., 1939, *Proc. Roy. Soc.* A, **170**, 513.
GUGGENHEIMER, K. M., 1934, *J. Phys. Radium*, **5**, 253, 475.
HAFSTAD, L. R., and TELLER, E., 1938, *Phys. Rev.*, **54**, 681.
VON HALBAN, H., and KOWARSKI, L., 1938, *Nature Lond.*, **142**, 392.
HARDY, W. A., SILVEY, G., and TOWNES, C. H., 1952, *Phys. Rev.*, **85**, 494.
HARVEY, J. A., 1951, *Phys. Rev.*, **81**, 353.
HAXEL, O., JENSEN, J. H. D., SUESS, H. E., 1948, *Naturwissenschaften*, **35**, 375; 1949 *Phys. Rev.*, **75**, 1766; 1950, *Z. Phys.*, **128**, 295.
HILL, D. L., and WHEELER, J. A., 1953, *Phys. Rev.*, **89**, 1102.
HILL, R. D., 1949, *Phys. Rev.*, **76**, 998.
HITCHCOCK, A. J. M., 1953, *Phil. Mag.*, **44**, 766.
HÖCKER, K. H., 1952, *Z. Phys.*, **132**, 399.
HOPKINS, N. J., 1952, *Phys. Rev.*, **88**, 680.
HORIE, H., UMEZAWA, M., YAMAGUCHI, Y., YOSHIDA, S., 1951, *Prog. Theor. Phys.*, **6**, 254.
HORIE, H., and YOSHIDA, S., 1951, *Prog. Theor. Phys.*, **6**, 829.
HUGHES, D. J., SPATZ, W. D. B., and GOLDSTEIN, N., 1949, *Phys. Rev.*, **75**, 1781.
HUND, F., 1937, *Z. Phys.*, **105**, 202.
INGLIS, D. R., 1953, *Rev. Mod. Phys.*, **25**, 390.
JAHN, H. A., 1950, *Proc. Roy. Soc.* A, **201**, 516; 1951, *Ibid.*, **205**, 192.
JAHN, H. A., and VAN WIERINGEN, H., 1951, *Proc. Roy. Soc.* A, **209**, 502.
KEILSON, J., 1951, *Phys. Rev.*, **82**, 759.
KING, R. W., and PEASLEE, D. C., 1953, *Phys. Rev.*, **90**, 1001.
KLINKENBERG, P. F. A., 1952, *Rev. Mod. Phys.*, **24**, 63.
KOHMAN, T. P., 1948, *Phys. Rev.*, **73**, 16; 1952, *Ibid.*, **85**, 530.
KOWARSKI, L., 1950, *Phys. Rev.*, **78**, 477.
KURATH, D., 1950, *Phys. Rev.*, **80**, 98; 1952, *Ibid.*, **88**, 804.
LATHAM, R., 1947, *Proc. Phys. Soc.*, **59**, 979.
MAYER, M. G., 1948, *Phys. Rev.*, **74**, 235; 1949, *Ibid.*, **75**, 1969; 1950 a, *Ibid.*, **78**, 16, 1950 b, *Ibid.*, **78**, 22.
MAYER, M. G., MOSKOWSKI, S. A., and NORDHEIM, L. W., 1951, *Rev. Mod. Phys.*, **23**, 315.
MEITNER, L., 1952, *Ark. Fys.*, **4**, 383.
MESCHERYAKOV, C. R., 1945, *C.R. Acad. Sci., U.S.S.R.*, **48**, 555.
MIYAZAWA, H., 1951, *Prog. Theor. Phys.*, **6**, 263, 801.
MIZUSHIMA, M., and UMEZAWA, M., 1951, *Phys. Rev.*, **83**, 463; 1952, *Ibid.*, **85**, 37, *Ibid.*, **86**, 1055.
NORDHEIM, L. W., 1949, *Phys. Rev.*, **75**, 1894; 1950, *Ibid.*, **78**, 294; 1951, *Rev. Mod. Phys.*, **23**, 322.
OSBORN, R. K., and FOLDY, L. L., 1950, *Phys. Rev.*, **79**, 1950.

PERLMAN, I., GHIORSO, A., and SEABORG, G. T., 1948, *Phys. Rev.*, **75**, 1730; 1950, *Ibid.*, **77**, 26.
PFIRSCH, D., 1952, *Z. Phys.*, **132**, 409.
PRYCE, M. H. L., 1950, *Proc. Phys. Soc.* A, **63**, 692; 1952, *Ibid.*, **65**, 773, corrigendum, 962.
RACAH, G., 1949, *Phys. Rev.*, **76**, 1352.
RAINWATER, J., 1950, *Phys. Rev.*, **79**, 432.
ROSENFELD, L., 1948, *Nuclear Forces* (Amsterdam: North Holland Publishing Co.); 1951, *Physica*, **17**, 461.
ROSS, M., 1952, *Phys. Rev.*, **88**, 935.
SACHS, R. G., 1948, *Phys. Rev.*, **74**, 433.
SCHARFF-GOLDHABER, G., 1953, *Phys. Rev.*, **90**, 587.
SCHMIDT, T., 1937, *Z. Phys.*, **106**, 358.
SCHWARTZ, H. M., 1953, *Phys. Rev.*, **89**, 1293.
DE SHALIT, A., 1951, *Helv. Phys. Acta*, **24**, 296; 1953, *Phys. Rev.*, **90**, 83.
DE SHALIT, A., and GOLDHABER, M., 1953, *Phys. Rev.*, **92**, 1211.
SPRUCH, L., 1950, *Phys. Rev.*, **80**, 372.
STÄHELIN, P., and PREISWERK, P., 1951, *Helv. Phys. Acta*, **24**, 623.
SWIATECKI, W. J., 1951, *Proc. Roy. Soc.* A, **205**, 238.
TALMI, I., 1951, *Phys. Rev.*, **82**, 101; 1952, *Helv. Phys. Acta*, **25**, 185.
TOWNES, C. H., FOLEY, H. M., and LOW, W., 1949, *Phys. Rev.*, **76**, 1415.
TROCHERIS, M., 1952, *J. Phys. Radium*, **13**, 299.
UMEZAWA, M., 1952, *Prog. Theor. Phys.*, **8**, 509.
VAN WAGENINGEN, R., and DE BOER, J., 1952, *Physica*, **18**, 369.
WAPSTRA, A. H., 1951, *Physica*, **17**, 628; 1952, *Ibid.*, **18**, 83.
WAY, K., and WOOD, M., 1951, *Neutron and Proton Binding Energy Map* (Washington D.C.: National Bureau of Standards).
WEFELMEIER, W., 1937, *Z. Phys.*, **107**, 332.
WEIZSÄCKER, C. F., 1935, *Z. Phys.*, **96**, 431.
WIGNER, E. P., 1937 a, *Phys. Rev.*, **51**, 106; 1937 b, *Ibid.*, **51**, 947; 1951, *University of Wisconsin Summer School*, Lecture Notes.

DIFFRACTION THEORY

By C. J. BOUWKAMP
Philips Research Laboratories, N. V. Philips' Gloeilampenfabrieken, Eindhoven, Netherlands

CONTENTS

PAGE
§ 1. Introduction, scope and limitation of the report 35
§ 2. General remarks on scalar diffraction problems for plane obstacles 38
§ 3. General remarks on electromagnetic diffraction problems for plane obstacles 43
§ 4. Kirchhoff's scalar theory of diffraction and its various modifications 49
§ 5. Approximations in electromagnetic diffraction theory 56
§ 6. Variational formulation of diffraction problems 62
§ 7. The Fourier transform method of solving integral equations of diffraction theory 66
§ 8. Scalar diffraction by small apertures and narrow slits 71
§ 9. Diffraction of a plane electromagnetic wave by a small circular obstacle 75
§ 10. Mathieu functions and spheroidal wave functions in diffraction theory .. 80
§ 11. Diffraction by spherical and cylindrical obstacles 85
§ 12. Survey of other problems in diffraction theory 90
References 92

Abstract. A critical review is presented of recent progress in classical diffraction theory. Both scalar and electromagnetic problems are discussed. The report may serve as an introduction to general diffraction theory although the main emphasis is on diffraction by plane obstacles. Various modifications of the Kirchhoff and Kottler theories are presented. Diffraction by obstacles small compared with the wavelength is discussed in some detail. Other topics included are: variational formulation of diffraction problems, the Wiener–Hopf technique of solving integral equations of diffraction theory, the rigorous formulation of Babinet's principle, the nature of field singularities at sharp edges, the application of Mathieu functions and spheroidal wave functions to diffraction theory. Reference is made to more than 500 papers published since 1940.

§ 1. Introduction, Scope and Limitation of the Report

A glance through the scientific literature of the last decade reveals an increasing interest in diffraction theory. Undoubtedly the impetus to new developments came not so much from classical optics as from radio microwave technique. Of course both fields are founded on Maxwell's equations. However, in most applications of diffraction theory to optics, light is considered as a scalar wave phenomenon, so that polarization is ignored. For example, calculations on diffraction of light by apertures are usually based on Kirchhoff's mathematical formulation of Huygens' principle. Experiments have shown that this is justifiable when the wavelength is small in comparison with the size of the aperture. Generally, for the purposes of instrumental optics the scalar theory is quite adequate (Theimer, Wassermann and Wolf 1952).

On the other hand, specific polarization effects cannot be ignored in radio-wave propagation when the wavelength is of the same order of magnitude as the aperture. One way to cover this case is by using a one-component Hertzian vector as the scalar wave function. Alternatively we may use Kottler's electromagnetic equivalent of Huygens' principle. If the scalar Kirchhoff formula is applied to each of the six rectangular components of the electric and magnetic vectors, then the six wave functions so obtained do not in general satisfy Maxwell's equations. This difficulty can be solved by introducing certain additional contour integrals, representing the effect of fictitious line charges along the rim of the aperture.

The theories of Kirchhoff and Kottler are poor substitutes for 'rigorous' diffraction theory (wave equation plus boundary conditions) in the quasi-optical range because they do not correctly describe the field in the vicinity of the aperture and the edge. In the extreme case of very long waves they entirely fail to predict the correct order of magnitude of the field far from the aperture (Rayleigh, see Bethe 1944).

Many attempts have been made to bridge the gap between the domains of applicability of the Kirchhoff–Kottler theory on the one hand and Rayleigh's first-order theory on the other hand. Various modifications of the Kirchhoff theory have been suggested, and higher-order approximations in Rayleigh's sense have been obtained for the slit and circular-hole problems. Fortunately several diffraction problems have been solved rigorously. For example, correct solutions are now available of the problems of diffraction of a plane polarized electromagnetic wave by a circular disc and by a circular hole in an infinite perfectly conducting screen. Except for the classical case of the sphere, these are the only essentially three-dimensional electromagnetic diffraction problems which have been solved so far. Further, the integral-equation approach to diffraction problems, originally suggested by Poincaré, was widely applied in recent years. The so-called Wiener–Hopf technique proved its power in the solution of a class of integral equations underlying certain waveguide problems. The rigorous form of Babinet's principle in electromagnetic diffraction has been found only recently, and new insight was gained into the nature of the field singularities at a sharp edge. Finally, variational techniques have shown their usefulness in a great number of diffraction problems.

The aim of this Report is to give an account of recent progress in the theory of diffraction as touched upon in the preceding paragraph. The main emphasis will be on planar diffraction theory from the point of view of wave equation plus boundary conditions. This includes the diffraction of guided waves. Both scalar and electromagnetic problems will be discussed. Apart from their intrinsic interest and applications to acoustics, scalar problems are of importance for electromagnetic theory because in many cases the results of scalar theory can at once be interpreted in terms of electromagnetic theory (e.g. two-dimensional diffraction problems). In addition, recent success in the solution of electromagnetic problems is largely due to the fact that similar scalar problems have been solved before.

It goes without saying that the field of scattering, reflection and diffraction of electromagnetic waves is much too large to be covered in one review by one author. For example, numerical results for the diffraction of an electromagnetic

wave by a spherical particle, although very useful in a great number of problems occurring in physics, chemistry, meteorology and radio engineering, are not included in this Report. For applications of diffraction theory to colloid chemistry the reader is referred to an exhaustive review by Oster (1948). Other applications are concerned with optics of opal glasses, scattering of light and radio waves by dust particles and water droplets in the atmosphere, and extinction of light by solid particles in interstellar space (Van de Hulst 1949b). It was for the latter purpose that Van de Hulst (1946) collected and orderly arranged all relevant computations on electromagnetic diffraction by spherical particles. The present author has been informed that a monograph on the subject by Van de Hulst is in preparation.

Another subject of diffraction theory that is fully excluded from this Report is the theory of radio-wave propagation around the spherical earth. In this connection the reader should consult a recent book by Bremmer (1949) or the exhaustive report of Burrows and Attwood (1949).

It should be noted that a number of review papers bearing on the general subject of the present Report have been published. For a survey of general laws in optical diffraction theory we refer to a critical review by Toraldo di Francia (1949, 1950b). Diffraction theory of optical instruments and the image formation in the presence of aberrations have been covered in a preceding volume of these Reports (Wolf 1951). Schoch (1950) has given a detailed account of acoustical aspects of diffraction theory. Recent progress in diffraction theory was critically reviewed by the author in a series of lectures given at New York University. These lectures have been published in the form of an E.M. report (Bouwkamp 1953). Various mathematical aspects of electromagnetic diffraction theory will be treated in a forthcoming review by Claus Müller.

It is worth while to mention that a second and revised edition of Baker and Copson's standard work on *The Mathematical Theory of Huygens' Principle* (Baker and Copson 1950) has been met with great enthusiasm by workers in the field. For an introduction in diffraction theory the reader may also consult the relevant chapters in Volume 4 of Sommerfeld's published lectures on theoretical physics (Sommerfeld 1950). Another important book, dealing with the engineering aspects of diffraction of guided waves by irises and similar obstacles in waveguides, is due to Marcuvitz (1951a). An extension of this work by Marcuvitz and Schwinger is planned for the near future

It should be mentioned that the present Report is concerned with classical diffraction theory only. Problems that are of primary importance in quantum-mechanical scattering processes have been omitted.

Only steady-state problems will be discussed. It is tacitly understood that the time factor is $\exp(-i\omega t)$, where i is the imaginary unit, ω the angular frequency, and t the time. The wave number is denoted by $k=2\pi/\lambda=\omega/c$, where c is the phase velocity in free space of plane waves, and λ the wavelength. Throughout the Report x, y, z denote right-handed Cartesian coordinates. Outside the sources of radiation and the diffracting obstacles, space is assumed to be homogeneous, isotropic, and non-absorbing ($k>0$). However, in connection with conditions at infinity it will sometimes be convenient to allow k to be complex, with the imaginary part small and positive. In electromagnetic diffraction problems we shall use the gaussian system of units.

§ 2. General Remarks on Scalar Diffraction Problems for Plane Obstacles

Consider the diffraction of an arbitrary incident wave $u_0(x, y, z)$ through an aperture A in an infinite plane screen S of vanishing thickness. Throughout the Report it will be assumed that the plane A + S coincides with the plane $z = 0$, and that the primary field u_0, which is a solution of the wave equation, is incident from the left ($z < 0$).

Two different problems can be distinguished according as the wave function or its normal derivative vanishes on S. In the first case, that is, for a perfectly soft screen, we have for the total field

$$u_{s1}(x, y, z) = \begin{cases} u_0(x, y, z) - u_0(x, y, -z) + \phi_1(x, y, -z) & (z \leqslant 0) \\ \phi_1(x, y, z) & (z \geqslant 0) \end{cases} \quad \ldots\ldots (2.1)$$

where ϕ_1, to be defined for $z \geqslant 0$ only, has the following properties:

(i) ϕ_1 is a solution of the wave equation, $\Delta\phi_1 + k^2\phi_1 = 0$, when $z > 0$;

(ii) $\phi_1 = 0$ on S;

(iii) ϕ_1 satisfies Sommerfeld's radiation condition at infinity (cf. Atkinson 1949, Svešnikov 1950);

(iv) $\partial\phi_1/\partial z = \partial u_0/\partial z$ in A;

(v) ϕ_1 is everywhere finite;

(vi) $\nabla\phi_1$ is quadratically integrable over any domain of three-dimensional space, including the edge of the screen.

It should be remarked that the first four properties are not sufficient to characterize the function ϕ_1 uniquely; other functions may be found by differentiation (Rayleigh, see Bouwkamp 1946 b). However, it may be conjectured that properties (i) to (v) inclusive do characterize ϕ_1 completely, in the sense that (vi) would then necessarily follow. In fact, all known solutions satisfy (vi), ensuring that no energy is radiated by the singularities occurring at the edge. It should also be noted that we cannot require quadratic integrability of $\nabla\phi_1$ over the *two*-dimensional domain of the aperture. We can, however, require absolute integrability of $\nabla\phi_1$ over the aperture.

The second boundary-value problem, in which the normal derivative of the wave function vanishes on the screen, applies to acoustic diffraction through an aperture in a perfectly rigid screen. In this case the total field is given by

$$u_{s2}(x, y, z) = \begin{cases} u_0(x, y, z) + u_0(x, y, -z) - \phi_2(x, y, -z) & (z \leqslant 0) \\ \phi_2(x, y, z) & (z \geqslant 0) \end{cases} \quad \ldots\ldots (2.2)$$

where ϕ_2, also defined for $z \geqslant 0$ only, has similar properties to ϕ_1 except that (ii) and (iv) should be replaced by

(ii)′ $\partial\phi_2/\partial z = 0$ on S;

(iv)′ $\phi_2 = u_0$ in A.

The remarks made on ϕ_1 concerning uniqueness and integrability apply equally to ϕ_2. There are good reasons to believe that ϕ_1 vanishes as $D^{1/2}$, and that $\nabla\phi_1$ and $\nabla\phi_2$ become infinitely large as $D^{-1/2}$, when the field point in free space approaches the edge of the screen; here D denotes the distance from the edge (cf. Bouwkamp 1946 b, Meixner 1949 b).

We now consider the complementary problem, that is, the diffraction by a disc of the same size and shape as the aperture A. Surprisingly, the two

corresponding boundary-value problems for the disc can be solved in terms of the very functions ϕ_1 and ϕ_2. For the perfectly soft disc we have

$$u_{d1}(x,y,z)=u_0(x,y,z)-\phi_2(x,y,\mp z) \qquad (z\lesssim 0), \qquad \ldots\ldots(2.3)$$

and for the perfectly rigid disc we have

$$u_{d2}(x,y,z)=u_0(x,y,z)\pm\phi_1(x,y,\mp z) \qquad (z\lesssim 0). \qquad \ldots\ldots(2.4)$$

As a consequence, the diffraction by a perfectly soft (rigid) disc is essentially identical with the diffraction by the complementary aperture in a perfectly rigid (soft) screen. More precisely, the rigorous form of Babinet's principle in acoustic diffraction theory (Bouwkamp 1941, Sommerfeld 1950, p. 277) is contained in the equations

$$u_{a1}\mp u_{d2}=u_{d1}\mp u_{a2}=\mp u_0(x,y,\mp z) \qquad (z\lesssim 0), \qquad \ldots\ldots(2.5)$$

where it should be recalled that the subscripts 'a' and 'd' refer to diffraction by the aperture and the complementary disc respectively, and the subscripts 1 and 2 refer to the type of boundary condition at the obstacle (vanishing u and $\partial u/\partial z$ respectively).

It is difficult, if possible at all, to determine the wave functions $\phi_{1,2}$ for an aperture of arbitrary shape. The separation technique is applicable only to an elliptical aperture, including the special cases of a circular aperture (Bouwkamp 1941, Kotani 1933, Spence 1949, Storruste 1948, Storruste and Wergeland 1948 a, b) and an infinite slit (Morse and Rubenstein 1938, Strutt 1931, Skavlem 1951). By this technique infinite-series expansions are obtained for $\phi_{1,2}$ in terms of normal solutions, that is, eigenfunctions of the wave equation after separation in elliptical coordinates. Numerical evaluation of the series is cumbersome and demands extensive auxiliary tables of Mathieu functions and their generalizations such as spheroidal wave functions. At present such tables are not available to a sufficient extent and accordingly numerical results available for the diffraction problems under consideration are restricted to plane-wave excitation where three times the wavelength is greater than the width of the slit or the diameter of the circular aperture. In addition, numerical computations were usually performed for the simplest possible case, that is, for normal incidence; details will be given later.

A much more powerful approach to the solution of our diffraction problems is by means of integral equations. As is well known, any solution u of the wave equation that is regular inside a closed surface Σ is expressible in terms of the boundary values on Σ of either u or $\partial u/\partial n$, at least if we know the corresponding Green's functions of the first and second kinds. The latter functions, however, are known in the case that Σ consists of an infinite plane and a half-sphere at infinity (cf. Foldy and Primakoff 1945). Accordingly, if u is any wave function that is regular for $z\geqslant 0$ and satisfies appropriate conditions at infinity (including the radiation condition), we have Rayleigh's often-quoted formulae

$$u=-\frac{1}{2\pi}\int\frac{\partial u}{\partial n}\frac{e^{ikr}}{r}\,d\Sigma, \qquad \ldots\ldots(2.6)$$

$$u=\frac{1}{2\pi}\int u\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)d\Sigma, \qquad \ldots\ldots(2.7)$$

where the integration is over the plane $z'=0$ and r denotes the distance between the field point (x,y,z) and the source point (x',y',z'), while $\partial/\partial n$ means differentiation with respect to z' and then putting $z'=0$.

In spite of the singularities of $\nabla\phi_{1,2}$ at the edge, it is legitimate to apply Rayleigh's formulae to our functions $\phi_{1,2}$. In virtue of (ii) and (ii)′ it is most convenient to apply eqn. (2.6) to ϕ_2 and eqn. (2.7) to ϕ_1 because then the screen does not contribute to the integrals. We thus find

$$\phi_1 = \frac{1}{2\pi}\int_A \phi_1 \frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right) d\Sigma, \qquad \ldots\ldots(2.8)$$

$$\phi_2 = -\frac{1}{2\pi}\int_A \frac{e^{ikr}}{r}\frac{\partial \phi_2}{\partial n}\, d\Sigma. \qquad \ldots\ldots(2.9)$$

For any shape of aperture we have now expressed $\phi_{1,2}$ in terms of the values of ϕ_1 and $\partial\phi_2/\partial z$ in the aperture. On applying the crucial relations (iv) and (iv)′ we get 'integral equations' for the unknown aperture values, which means that our problem is reduced to one in two dimensions. The case of ϕ_2 is simple. Let us introduce the kernel function

$$G(x,y;x',y') = \frac{\exp[ik\{(x-x')^2+(y-y')^2\}^{1/2}]}{\{(x-x')^2+(y-y')^2\}^{1/2}}.$$

Then (2.9) and (iv)′ lead to the pure integral equation

$$\int_A \left\{\frac{\partial}{\partial z'}\phi_2(x',y',0)\right\} G(x,y;x',y')\, dx'\, dy' = -2\pi u_0(x,y,0), \qquad \ldots\ldots(2.10)$$

where $(x,y,0)$ is any point of the aperture. In the case of ϕ_1 many authors (e.g. Levine and Schwinger 1948 b) on applying (iv) to (2.8) differentiate with respect to z under the sign of integration in (2.8), so as to obtain formally a pure integral equation with the (non-integrable!) kernel

$$\left[\frac{\partial^2}{\partial z \partial z'}\left(\frac{e^{ikr}}{r}\right)\right]_{z=z'=0}.$$

This procedure eventually leads to all kinds of divergent integrals (Baker and Copson 1950, p. 186) and should therefore be rejected. The divergencies are easy to eliminate (Bouwkamp 1953) and it turns out that a differential-integral equation, rather than an integral equation, is obtained, viz.

$$\left[k^2+\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right]\int_A G(x,y;x',y')\phi_1(x',y',0)\, dx'\, dy' = 2\pi\frac{\partial}{\partial z}u_0(x,y,0), \qquad .(2.11)$$

where $(x,y,0)$ is in the aperture A (cf. Baker and Copson 1950, p. 184). Maue (1949 b, eqn. (60)) has given an equivalent of eqn. (2.11), namely,

$$k^2\int \phi_1 G\, d\Sigma - \int \nabla\phi_1 . \nabla G\, d\Sigma = 2\pi\frac{\partial u_0}{\partial z}, \qquad \ldots\ldots(2.12)$$

where both gradients are to be taken with respect to the source-point coordinates. Maue's equation follows from (2.11) by a process of differentiation and integration by parts, and use of the condition $\phi_1 = 0$ at the edge. The second integral in (2.12) is a Cauchy's principal value.

Only in a few simple cases has the differential-integral equation (2.11) been transformed into a pure integral equation. If the incident field is a plane wave,

$$u_0 = \exp[ik(\alpha_0 x+\beta_0 y+\gamma_0 z)], \qquad \gamma_0 = (1-\alpha_0^2-\beta_0^2)^{1/2},$$

we get

$$\int_A G(x,y;x',y')\phi_1(x',y',0)\, dx'\, dy' = (2\pi i/k\gamma_0)\exp[ik(\alpha_0 x+\beta_0 y)] + \chi(x,y), \qquad \ldots\ldots(2.13)$$

where χ is a solution of the two-dimensional wave equation in the aperture.

The function χ can be uniquely determined (except for a constant multiplier) for a normally incident wave in the cases of the circular aperture (Levine and Schwinger 1948 b) and the infinite slit (Sommerfeld 1950, p. 283). The resulting equation can be transformed into a pure integral equation with a non-symmetrical kernel.

Other types of integral equation can be obtained if eqn. (2.6) is applied to ϕ_1 and eqn. (2.7) to ϕ_2 and use is made of (iv) and (iv)'. Then a pure integral equation is found in the first case and a differential-integral equation in the second case. Confining ourselves to ϕ_1, we have (Levine and Schwinger 1949 a)

$$\int_S G(x,y;x',y')\left\{\frac{\partial}{\partial z'}\phi_1(x',y',0)\right\}dx'\,dy' = -\int_A G(x,y;x',y')\left\{\frac{\partial}{\partial z'}u_0(x',y',0)\right\}dx'\,dy', \quad \ldots\ldots(2.14)$$

where the unknown function is $\partial\phi_1/\partial z$ on S while the right-hand side of eqn. (2.14) may be considered known. An equivalent, but simpler form of (2.14) was given by Baker and Copson (1950, p. 164). The function u_{a1} is everywhere given by

$$u_{a1} = u_0 - \frac{1}{2\pi}\int_S \left(\frac{\partial\phi_1}{\partial n} - \frac{\partial u_0}{\partial n}\right)\frac{e^{ikr}}{r}\,d\Sigma.$$

Therefore it suffices to know the values of $(\partial\phi_1/\partial n - \partial u_0/\partial n)$ on the screen S. Let these values be $h(x,y)$. Since u_{a1} must vanish on S we find

$$\int_S G(x,y;x',y')h(x',y')\,dx'\,dy' = 2\pi u_0(x,y,0), \quad \ldots\ldots(2.15)$$

where $(x,y,0)$ is on S. The integral equations (2.14) and (2.15) are not very suitable for calculations on small apertures because the domain S is unbounded. In this respect eqns. (2.10) and (2.11) are to be preferred (Baker and Copson 1950, p. 183).

A simple approach to the solution of diffraction problems, based on the representation of the scattered field as an angular spectrum of plane waves, was suggested by several authors independently (Booker and Clemmov 1950 a, Clemmov 1951, Durand 1948 c, Hönl 1952, Scheffers 1942, Silver 1949, Toraldo di Francia 1949). This representation shows that polar diagrams and aperture distributions are related by two-dimensional Fourier transformation. Further, our boundary-value problems can then be formulated in terms of a pair of dual integral equations (Clemmov 1951, Hönl 1952).

To show the connection with the earlier results, we expand the free-space Green's function into a system of plane waves propagating in all directions, including complex ones (cf. Weyl 1919):

$$\frac{e^{ikr}}{r} = \frac{ik}{2\pi}\int_\Omega \exp\left[ik\{\alpha(x-x')+\beta(y-y')+\gamma|z-z'|\}\right]d\Omega, \quad \ldots\ldots(2.16)$$

where $\alpha = \sin\vartheta\cos\phi$, $\beta = \sin\vartheta\sin\phi$, $\gamma = \cos\vartheta$ are the directional cosines, and $d\Omega = \sin\vartheta\,d\vartheta\,d\phi$ is the surface element of the complex unit sphere Ω; the variables of integration ϕ and ϑ run from 0 to 2π and from 0 to $\frac{1}{2}\pi - i\infty$ respectively. If eqn. (2.16) is substituted in eqns. (2.8) and (2.9) it is readily verified that

$$\phi_{1,2} = \frac{ik}{2\pi}\int_\Omega A_{1,2}(\alpha,\beta)\exp\{ik(\alpha x+\beta y+\gamma z)\}\,d\Omega, \quad \ldots\ldots(2.17)$$

or, equivalently,

$$\phi_{1,2} = \frac{ik}{2\pi} \iint_{-\infty}^{\infty} A_{1,2}(\alpha, \beta) \exp\{ik(\alpha x + \beta y + \gamma z)\}\, d\alpha\, d\beta/\gamma, \quad \ldots\ldots(2.18)$$

where the angular-spectrum densities are defined by

$$A_1(\alpha, \beta) = -\frac{ik\gamma}{2\pi} \int_A \exp\{-ik(\alpha x' + \beta y')\} \phi_1(x', y', 0)\, dx'\, dy', \quad \ldots\ldots(2.19)$$

$$A_2(\alpha, \beta) = -\frac{1}{2\pi} \int_A \exp\{-ik(\alpha x' + \beta y')\} \frac{\partial}{\partial z'} \phi_2(x', y', 0)\, dx'\, dy', \quad \ldots\ldots(2.20)$$

respectively. On the other hand, we know that at large distances R from the aperture the diffracted wave becomes approximately spherical. The coefficient of $R^{-1}\exp(ikR)$ in the asymptotic expression for ϕ is called the amplitude $A(\alpha, \beta)$, which determines the polar diagram as a function of the directional cosines α and β (with respect to the x and y axes). By expanding the exponential function in eqns. (2.8) and (2.9) and observing that $r \simeq R - (\alpha x' + \beta y')$, we at once find that the amplitudes are simply identical with those defined by (2.19) and (2.20). Comparing eqn. (2.18) for $z = 0$ with eqn. (2.19) we see that the aperture distribution $\phi_1(x, y, 0)$ and the (modified) amplitude $(i/\gamma)A_1(\alpha, \beta)$ are each other's Fourier transforms, and so are $(\partial/\partial z)\phi_2(x, y, 0)$ and $-kA_2(\alpha, \beta)$ in the second case. Note that $\gamma = (1 - \alpha^2 - \beta^2)^{1/2}$ for the homogeneous, and $\gamma = i(\alpha^2 + \beta^2 - 1)^{1/2}$ for the inhomogeneous, plane-wave components (the latter correspond to surface waves or evanescent waves, erroneously omitted by Booker and Clemmov in their paper cited above). The dual integral equations of Clemmov (1951) and Hönl (1952) are obtained from eqn. (2.17) by putting $z = 0$ and applying conditions (ii) and (iv) or (ii)′ and (iv)′. For circular apertures and axial symmetry these equations can be written in terms of Bessel functions (see for example King 1934).

To calculate the energy transmitted by the aperture we integrate the square of the absolute value of the amplitude over all real directions in the half-space $\alpha, \beta \geqslant 0$. An alternative way is to evaluate the energy flux through the aperture more directly by noting (Levine and Schwinger 1948 b) that $(1/2ik)(\phi^* \nabla\phi - \phi \nabla\phi^*)$ represents the energy flux density (for a simple plane wave of unit amplitude it is the unit vector in the direction of propagation). Generally we have the connecting relation

$$-\frac{1}{ik} \int_A \phi_{1,2} \frac{\partial \phi_{1,2}^*}{\partial n}\, d\Sigma = \int_\Omega |A_{1,2}|^2\, d\Omega, \quad \ldots\ldots(2.21)$$

which is similar to Parseval's theorem in Fourier analysis and easy to prove from eqn. (2.17). The real part of either side of eqn. (2.21) is the transmission cross section of the aperture, that is, the ratio of the total transmitted power to that falling on the aperture per unit area normal to the direction of the incident wave. In virtue of eqn. (2.21) the reactive part of the cross section can be calculated also from the polar diagram if the latter is continued analytically into complex directions (cf. Bouwkamp 1946 a). According to a remarkable theorem by Levine and Schwinger (1948 b) the plane-wave transmission cross section of the aperture is simply proportional to the imaginary part of the amplitude of the scattered spherical wave observed in the direction of incidence of the incoming plane wave, viz.

$$\sigma_{1,2}(\alpha_0, \beta_0) = -\frac{2\pi}{k} \mathcal{I}\{A_{1,2}(\alpha_0, \beta_0)\}, \quad \ldots\ldots(2.22)$$

where the incident wave is $u_0 = \exp[ik(\alpha_0 x + \beta_0 y + \gamma_0 z)]$. To prove this, we take the real part of the left-hand side of eqn. (2.21), eliminate one of the factors in the integrand by using (iv) or (iv)′ and compare the resulting expressions with the definitions (2.19) and (2.20). A relation of the form (2.22) holds in many scattering problems (Feenberg 1932, Van de Hulst 1949a, Lax 1950).

In concluding this section we consider the integral-equation formulation for two-dimensional diffraction problems. It will be assumed that the various wave functions are independent of the x coordinate. The relevant formulae are readily obtained from their analogues in three dimensions if it is remembered that

$$\int_{-\infty}^{\infty} \frac{e^{ikr}}{r}\, dx' = \pi i H_0^{(1)}(k\rho)$$

where $r^2 = (x-x')^2 + (y-y')^2 + (z-z')^2$, $\rho^2 = (y-y')^2 + (z-z')^2$, and $H_0^{(1)}$ is the Hankel function of the first kind and of order zero. For example, eqns. (2.8) and (2.9) now become

$$\phi_1(y,z) = -\frac{i}{2}\frac{\partial}{\partial z}\int \phi_1(y',0) H_0^{(1)}[k\{(y-y')^2 + z^2\}^{1/2}]\, dy', \qquad \ldots\ldots (2.23)$$

$$\phi_2(y,z) = -\frac{i}{2}\int \left\{\frac{\partial}{\partial z'}\,\phi_2(y',0)\right\} H_0^{(1)}[k\{(y-y')^2 + z^2\}^{1/2}]\, dy', \qquad \ldots\ldots (2.24)$$

where the integration is over the intersection of the aperture with the plane $x = 0$. If the aperture is a simple slit of width $2b$, eqns. (2.10) and (2.11) reduce to

$$\int_{-b}^{b} H_0^{(1)}(k|y-y'|) \left\{\frac{\partial}{\partial z'}\,\phi_2(y',0)\right\} dy' = 2iu_0(y,0), \qquad \ldots\ldots (2.25)$$

$$\left(k^2 + \frac{d^2}{dy^2}\right)\int_{-b}^{b} H_0^{(1)}(k|y-y'|)\phi_1(y',0)\, dy' = -2i\,\frac{\partial}{\partial z}\,u_0(y,0), \qquad \ldots\ldots (2.26)$$

in which $-b \leqslant y \leqslant b$. Small errors in Sommerfeld's formulae (Sommerfeld 1950) have been observed by Groschwitz and Hönl (1952) to the effect that in Sommerfeld's book, pp. 284–295, the constants A and A' are to be replaced by $(2/\pi i)A$ and $(2/\pi i)A'$.

For two-dimensional diffraction problems eqn. (2.22) remains true if the amplitude of the cylindrical wave at large distances ρ from the aperture is defined to be the coefficient of $(2\pi/k\rho)^{1/2}\exp[i(k\rho + \pi/4)]$ in the asymptotic expression for $\phi_{1,2}$.

§ 3. General Remarks on Electromagnetic Diffraction Problems for Plane Obstacles

For steady-state phenomena in free space, Maxwell's equations reduce to

$$\operatorname{curl}\mathbf{E} = ik\mathbf{H}, \qquad \operatorname{curl}\mathbf{H} = -ik\mathbf{E}, \qquad \ldots\ldots (3.1)$$

so that both $\mathbf{E}$ and $\mathbf{H}$ satisfy the time-independent *vector wave equation*

$$\operatorname{curl}\operatorname{curl}\mathbf{E} - k^2\mathbf{E} = \operatorname{curl}\operatorname{curl}\mathbf{H} - k^2\mathbf{H} = 0. \qquad \ldots\ldots (3.2)$$

If $\mathbf{E}$ and $\mathbf{H}$ are resolved in rectangular Cartesian components, any of the six scalar functions so obtained satisfies the *scalar wave equation*. However, this property is of little value if we are concerned with diffraction problems for obstacles of arbitrary shape because then the boundary conditions for the rectangular components become much too involved, even for perfectly conducting obstacles. Further, separation of the vector wave equation in the

standard curvilinear coordinate systems is not always possible (C. Müller 1949a, b, Spencer 1951). By way of illustration it may be mentioned that so far all attempts (Herzfeld 1911, Möglich 1927) to solve the problem of the diffraction of a plane electromagnetic wave by a perfectly conducting ellipsoid have been without success. Only the simplest case has been conquered (Meixner 1948b, Meixner and Andrejewski 1950, Andrejewski 1953), that is, for an oblate spheroid of negligible thickness (circular disc).

As mentioned in §1, two-dimensional diffraction problems are essentially scalar. If the obstacle is cylindrical in the direction of the x axis and if, moreover, the incident plane wave is polarized either parallel or perpendicular to the x axis, then the field configuration is independent of x. In the first case (parallel polarization) the electric field has only one component, E_x, and in the second case (perpendicular polarization) there is only one magnetic component, H_x. In virtue of eqns. (3.1) the remaining non-zero components follow from E_x or H_x by differentiation, so that in both cases the total field can be expressed in terms of one scalar wave function and its derivatives.

For plane obstacles (half-plane, strip, grating, etc.) of infinite conductivity and negligible thickness the explicit expressions for the field components are, in an obvious notation,

$$\mathbf{E}=(u_1,0,0), \quad ik\mathbf{H}=\left(0,\frac{\partial u_1}{\partial z},-\frac{\partial u_1}{\partial y}\right) \qquad \ldots\ldots(3.3)$$

if the incident field is polarized parallel to the edge, and

$$\mathbf{H}=(u_2,0,0), \quad -ik\mathbf{E}=\left(0,\frac{\partial u_2}{\partial z},-\frac{\partial u_2}{\partial y}\right) \qquad \ldots\ldots(3.4)$$

if the incident field is polarized perpendicular to the edge. Here u_1 and u_2 are identical with the scalar wave functions u_{d1} and u_{d2}, respectively, discussed in § 2.

For a proper insight into the nature of field singularities occurring at a sharp edge, we include in this Report the explicit expressions for the field components in Sommerfeld's half-plane diffraction problem (Bouwkamp 1946b). Let $z=0$, $y>0$ be the screen, and let us define polar coordinates ρ and θ by $y=\rho\cos\theta$, $z=\rho\sin\theta$. Further, let $u_0=\exp[ik\rho\cos(\theta-\theta_0)]$ be the incident plane wave where θ_0 is the angle between the positive y axis and the direction of incidence. Then

$$\left.\begin{matrix}u_1\\u_2\end{matrix}\right\}=\exp\{ik\rho\cos(\theta-\theta_0)\}\frac{\exp(-i\pi/4)}{\pi^{1/2}}\int_{-\infty}^{s_1}\exp(i\tau^2)\,d\tau$$
$$\mp\exp\{ik\rho\cos(\theta+\theta_0)\}\frac{\exp(-i\pi/4)}{\pi^{1/2}}\int_{-\infty}^{s_2}\exp(i\tau^2)\,d\tau, \qquad \ldots\ldots(3.5)$$

where $s_1=(2k\rho)^{1/2}\sin\frac{1}{2}(\theta-\theta_0)$, $s_2=-(2k\rho)^{1/2}\sin\frac{1}{2}(\theta+\theta_0)$. This result can be derived in many ways; a simple derivation by 'separation' of the wave equation in semi-parabolical coordinates, is implicit in a recent paper by J. Brillouin (1949), see Bouwkamp (1953). Using eqns. (3.3) to (3.5) we find, in the parallel case,

$$\left.\begin{aligned}&E_x=u_1,\\&H_y=u_2\sin\theta_0-\cos\tfrac{1}{2}\theta_0\cos\tfrac{1}{2}\theta\left(\frac{2}{\pi k\rho}\right)^{1/2}\exp\{i(k\rho+\pi/4)\},\\&H_z=-u_1\cos\theta_0-\cos\tfrac{1}{2}\theta_0\sin\tfrac{1}{2}\theta\left(\frac{2}{\pi k\rho}\right)^{1/2}\exp\{i(k\rho+\pi/4)\},\end{aligned}\right\} \qquad \ldots\ldots(3.6)$$

and in the perpendicular case,

$$\left.\begin{aligned} H_x &= u_2, \\ E_y &= -u_1 \sin\theta_0 - \sin\tfrac{1}{2}\theta_0 \sin\tfrac{1}{2}\theta \left(\frac{2}{\pi k\rho}\right)^{1/2} \exp\{i(k\rho+\pi/4)\}, \\ E_z &= u_2 \cos\theta_0 + \sin\tfrac{1}{2}\theta_0 \cos\tfrac{1}{2}\theta \left(\frac{2}{\pi k\rho}\right)^{1/2} \exp\{i(k\rho+\pi/4)\}. \end{aligned}\right\} \quad \ldots\ldots(3.7)$$

If the incident wave is polarized at an arbitrary angle with the edge we must take a linear combination of eqns. (3.6) and (3.7). The generalization of these equations to arbitrarily oblique incidence has been found only recently (Copson 1950, Clemmov 1951), while the earlier results of Wiegrefe (1913) show inadmissible singularities at the edge. Inspection of eqns. (3.6) and (3.7) and their generalizations by Copson and Clemmov reveals that the component of the electric field tangential to the edge vanishes at the edge as $D^{1/2}$, where D is the distance from the edge. The tangential component of the magnetic field remains also finite at the edge, but all other components of the electromagnetic field become infinitely large at the edge as $D^{-1/2}$. Despite these singularities, the edge does not radiate energy: the vectors **E** and **H** are quadratically integrable over any domain of three-dimensional space. It may be conjectured (Bouwkamp 1946b, 1950f) that the same types of singularity occur at any infinitely sharp edge, no matter whether the obstacle is plane or curved and independent of the form of the edge. A conclusive proof has not yet been given, since that would require a general existence theorem for the solution of the diffraction problem (under certain plausible and sufficient conditions). If it is assumed that the field components can be expanded into infinite series of a certain type, a proof of the theorem is possible (Meixner 1949b). In a somewhat different approach, using integral equations, emphasis is laid on the behaviour of the current and charge densities at the edge (Maue 1949b, Jones 1950, 1952a, b, Bouwkamp 1950f, 1953). It appears that the component of the current density normal to the edge vanishes at the edge as $D^{1/2}$ while the component tangential to the edge becomes infinitely large as $D^{-1/2}$. The charge density becomes infinitely large also as $D^{-1/2}$. In fact, these types of singularity occur in all known solutions. They seem to provide the 'natural' boundary conditions at a sharp edge. If these natural boundary conditions would always apply, it would be unnecessary to include additional contour integrals along the edge of the obstacle (as in the classical Huygens–Kottler formulation, where the contour integrals represent the effect of artificial line charges).

Let $(\mathbf{f}, \mathbf{g})$ denote any incident field, where $\mathbf{f}$ stands for the electric vector and $\mathbf{g}$ for the magnetic vector. The 'complementary incident field' is defined by $(-\mathbf{g}, \mathbf{f})$, in the order (electric, magnetic) vector. Both fields satisfy Maxwell's equations. First we consider the diffraction of the field $(\mathbf{f}, \mathbf{g})$ by a perfectly conducting plane screen S of zero thickness. Secondly we consider the diffraction of the complementary field $(-\mathbf{g}, \mathbf{f})$ by an aperture A in a perfectly conducting screen such that the aperture in the second problem is of the same size and shape as the screen in the first problem (A = S). For simplicity we call the second diffraction problem the 'complementary diffraction problem'. The rigorous form of Babinet's principle asserts that the solution of one of these problems gives at once the solution of the other problem (Copson 1946b, Meixner 1948b). In the first problem the total field everywhere in space is

given by $(\mathbf{f}+\mathbf{E}^s, \mathbf{g}+\mathbf{H}^s)$, where the scattered field $(\mathbf{E}^s, \mathbf{H}^s)$ is due to the electric currents induced in the screen by the incident field. In the complementary problem we can distinguish between the fields in front of and behind the aperture. Let $(\mathbf{E}_0, \mathbf{H}_0)$ denote the total field in the illuminated half-space $(z \leqslant 0)$ if there is no hole in the screen, and let $(\mathbf{E}^d, \mathbf{H}^d)$ denote the diffracted field in the presence of the aperture. The latter field constitutes the total field behind the aperture, but in front of the aperture the total field is $(\mathbf{E}_0+\mathbf{E}^d, \mathbf{H}_0+\mathbf{H}^d)$. Now according to the rigorous form of Babinet's principle we have the simple relations

$$\mathbf{E}^d = \mp \mathbf{H}^s, \qquad \mathbf{H}^d = \pm \mathbf{E}^s \qquad (z \lesseqgtr 0). \qquad \ldots\ldots(3.8)$$

It should be mentioned that the x and y components of $\mathbf{E}^{s,d}$ and the z components of $\mathbf{H}^{s,d}$ are symmetrical with respect to the plane $z=0$. On the other hand, the x and y components of $\mathbf{H}^{s,d}$ and the z components of $\mathbf{E}^{s,d}$ are odd functions of z. This is obvious, at least for the first problem, if we recall that $(\mathbf{E}^s, \mathbf{H}^s)$ is derivable from the surface current density $\mathbf{I}$ in the screen and the vector potential $\mathbf{A}^s$ by

$$\left.\begin{aligned} &\mathbf{A}^s = \frac{1}{c}\int_S \mathbf{I}\frac{e^{ikr}}{r}\,d\Sigma, \\ &\mathbf{H}^s = \operatorname{curl}\mathbf{A}^s, \qquad -ik\mathbf{E}^s = k^2\mathbf{A}^s + \operatorname{grad}\operatorname{div}\mathbf{A}^s. \end{aligned}\right\} \qquad \ldots\ldots(3.9)$$

In addition it follows that H_x^s, H_y^s and E_z^s vanish at $z=0$ outside the screen, which means that *in the aperture the tangential magnetic field and the normal electric field are not disturbed by the presence of the screen.*

In analogy with eqn. (3.9) the field $(\mathbf{E}^d, \mathbf{H}^d)$ may be derived from fictitious magnetic currents $\mathbf{K}$ in the aperture $A(=S)$ and the corresponding magnetic vector potential $\mathbf{F}^d$ by means of

$$\left.\begin{aligned} &\mathbf{F}^d = \pm\frac{1}{c}\int_A \mathbf{K}\frac{e^{ikr}}{r}\,d\Sigma, \qquad (z \lesseqgtr 0) \\ &\mathbf{E}^d = \operatorname{curl}\mathbf{F}^d, \qquad ik\mathbf{H}^d = k^2\mathbf{F}^d + \operatorname{grad}\operatorname{div}\mathbf{F}^d. \end{aligned}\right\} \qquad \ldots\ldots(3.10)$$

The electric current density $\mathbf{I}$ in the first problem and the magnetic current density $\mathbf{K}$ in the complementary problem are identical two-component vectors, $\mathbf{I}=\mathbf{K}$, and therefore the behaviour of $\mathbf{K}$ at the edge is the same as that of $\mathbf{I}$.

Meixner (1948 b), using the field-component representation and the various symmetry properties of the scattered and diffracted fields, has given a simple proof of eqns. (3.8). His earlier presentation of Babinet's principle was incorrect (Meixner 1946 b). A proof by Copson (1946 b), essentially based on scalar and vector potentials, requires qualification because of the edge conditions (Bouwkamp 1947 a). To avoid this difficulty it is convenient to eliminate the scalar potentials (Copson 1950) and to use eqns. (3.9) and (3.10) as a starting-point. Of other authors that have dealt with Babinet's principle in its rigorous form we mention Booker (1946), Leontovič (1946), Fel'd (1948), Maue (1949 a), Vassenr (1949 b, 1952 d), Aoki (1949 a), Aoki, Sunouchi and Utagawa (1949), Simon (1951), and Elliott (1952).

It may be noted that taking the complementary field (transformation $\mathbf{f}, \mathbf{g} \rightarrow -\mathbf{g}, \mathbf{f}$) of a plane wave is equivalent to rotating the plane of polarization through a right angle counter-clockwise looking in the direction of propagation. With this in mind, it is easy to verify that the two 'independent' sets of Sommerfeld's formulae (3.6) and (3.7) follow from each other by virtue of Babinet's principle!

Once we know the current density **I** our first problem may be considered to be solved. Relations for **I** in the form of differential-integral equations are obtained if we apply the well-known boundary conditions at the perfectly conducting screen, requiring that the tangential electric and normal magnetic components shall vanish at S. These relations are (Copson 1950)

$$\frac{\partial A_y^s}{\partial x} - \frac{\partial A_x^s}{\partial y} = -g_z, \quad k^2 A_x^s + \Delta A_x^s = \frac{\partial g_z}{\partial z}, \quad k^2 A_y^s + \Delta A_y^s = -\frac{\partial g_z}{\partial z}, \quad \ldots\ldots (3.11)$$

where $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2$. Equations (3.11) hold for all points $(x, y, 0)$ of S. They are differential-integral equations for the current density since

$$A_x = \frac{1}{c}\int_S I_x \frac{e^{ikr_0}}{r_0} d\Sigma, \qquad A_y = \frac{1}{c}\int_S I_y \frac{e^{ikr_0}}{r_0} d\Sigma, \qquad \ldots\ldots (3.12)$$

where $r_0 = [(x-x')^2 + (y-y')^2]^{1/2}$.

The tangential components of $\mathbf{H}^s$ and the normal component of $\mathbf{E}^s$ on the surface of the screen follow directly from

$$H_x^s = \pm \frac{2\pi}{c} I_y, \quad H_y^s = \mp \frac{2\pi}{c} I_x, \quad E_z^s = \pm 2\pi\sigma, \quad (z = \pm 0) \quad \ldots\ldots (3.13)$$

where σ is the surface charge density related to the currents by the equation of continuity

$$ikc\sigma = \operatorname{div} \mathbf{I} = \frac{\partial I_x}{\partial x} + \frac{\partial I_y}{\partial y}.$$

Further, a simple transformation with the use of Stokes' theorem leads to

$$\operatorname{div} \mathbf{A}^s = ik \int_S \sigma \frac{e^{ikr}}{r} d\Sigma + \frac{1}{c}\int_C I_n \frac{e^{ikr}}{r} ds, \qquad \ldots\ldots (3.14)$$

where the second integral is extended along the edge C of the screen and I_n is the projection of **I** on the inward normal of C. Because of the edge condition the contour integral vanishes, and we are left with

$$\mathbf{E}^s = ik\mathbf{A}^s - \operatorname{grad} \Psi, \qquad \Psi = \int_S \sigma \frac{e^{ikr}}{r} d\Sigma, \qquad \ldots\ldots (3.15)$$

where Ψ is the scalar potential of the surface charges σ. For a critical discussion of Copson's equations (3.11), or their equivalents in terms of vector *and* scalar potentials (Copson 1946 b), in connection with the singularities at the edge, the reader should consult section 8 of the author's report (Bouwkamp 1953).

As to the complementary problem, we remark that, so long as **K** is integrable, the field defined by eqns. (3.10) satisfies Maxwell's equations and the radiation condition in the half-spaces $z<0$ and $z>0$, and it satisfies the appropriate boundary conditions at the screen (vanishing tangential electric and normal magnetic fields). Further, no matter what **K** is, the tangential electric and normal magnetic field components are automatically continuous in the aperture. The only conditions that are not automatically satisfied are that the normal **E** and the tangential **H** components should be continuous in the aperture. These requirements lead to eqns. (3.11) with $\mathbf{A}^s$ replaced by $\mathbf{F}^d$ and **I** by **K**. At the same time it then turns out that E_z, H_x, H_y in the aperture *are identical with the corresponding undisturbed quantities of the incident field.* The analogues of eqns. (3.14) and (3.15) are ($z>0$)

$$\operatorname{div} \mathbf{F}^d = -ik \int_A \eta \frac{e^{ikr}}{r} d\Sigma - \frac{1}{c}\int_C K_n \frac{e^{ikr}}{r} ds, \qquad \ldots\ldots (3.16)$$

and

$$\mathbf{H}^d = -ik\mathbf{F}^d - \operatorname{grad} X, \qquad X = \int_A \eta \frac{e^{ikr}}{r} d\Sigma, \qquad \ldots\ldots (3.17)$$

where X is the scalar magnetic potential of the magnetic charges with density $\eta = (1/ikc)\,\mathrm{div}\,\mathbf{K}$. Vasseur (1949 a) does not use the natural edge condition $K_n = 0$ on C in his discussion of Babinet's principle.

A somewhat different approach in deriving eqns. (3.9) and (3.10) is as follows. The source-free field in the half-space $z > 0$ can be expressed in terms of the tangential components of the magnetic field on the plane $z = 0$. Let $\mathbf{n}$ denote the unit vector in the positive z direction. Then define a field $\mathbf{E}, \mathbf{H}$ by

$$\mathbf{H} = \mathrm{curl}\,\mathbf{A}, \qquad -ik\mathbf{E} = k^2\mathbf{A} + \mathrm{grad}\,\mathrm{div}\,\mathbf{A},$$

where the vector $\mathbf{A}$ is expressed as an integral over the (given) tangential magnetic field at $z = 0$ by

$$\mathbf{A} = \frac{1}{2\pi}\int (\mathbf{n} \times \mathfrak{H}) \frac{e^{ikr}}{r}\, d\Sigma. \qquad \ldots\ldots (3.18)$$

Since $H_x = -\partial A_y/\partial z$, $H_y = \partial A_x/\partial z$ it is easy to verify that the limiting values of H_x and H_y as $z \to +0$ are precisely the given values $\mathscr{H}_x$ and $\mathscr{H}_y$. The connection with eqn. (3.9) is obvious since $(c/2\pi)(\mathbf{n} \times \mathfrak{H})$ is just the current density on $z = +0$ (see eqns. (3.13)), so that $\mathbf{A}$ is in fact the vector potential. The corresponding form of eqn. (3.14) is

$$(ik\Psi =)\,\mathrm{div}\,\mathbf{A} = \frac{ik}{2\pi}\int_S (\mathbf{n}\,.\,\mathfrak{E}) \frac{e^{ikr}}{r}\, d\Sigma + \frac{1}{2\pi}\int_C (\mathbf{s}\,.\,\mathfrak{H}) \frac{e^{ikr}}{r}\, ds, \quad \ldots\ldots (3.19)$$

where $\mathbf{s}$ is the unit vector tangential to the edge C in the positive direction, and $\mathfrak{E}$ is the electric field on the screen. Given $\mathfrak{H}_{\tan}$, we know the value of $(\mathbf{n}\,.\,\mathfrak{E}) = \mathscr{E}_z = (1/ik)(\partial\mathscr{H}_x/\partial y - \partial\mathscr{H}_y/\partial x)$. For arbitrarily given tangential magnetic field on S and zero field outside S we must in general include the edge integral representing the contribution of a line charge of density $\tau = (1/2\pi ik)(\mathfrak{H}\,.\,\mathbf{s})$ to the scalar potential Ψ.

In the dual case, let the tangential electric field be given on $z = +0$. We then find for $z > 0$ the field

$$\mathbf{E} = \mathrm{curl}\,\mathbf{F}, \qquad ik\mathbf{H} = k^2\mathbf{F} + \mathrm{grad}\,\mathrm{div}\,\mathbf{F},$$

where

$$\mathbf{F} = \frac{1}{2\pi}\int (\mathbf{n} \times \mathfrak{E}) \frac{e^{ikr}}{r}\, d\Sigma \qquad \ldots\ldots (3.20)$$

is the magnetic vector potential of the magnetic currents of density $\mathbf{K} = -(c/2\pi)(\mathbf{n} \times \mathfrak{E})$ at $z = +0$. An easy transformation gives (Smythe 1947)

$$\mathbf{E} = \frac{1}{2\pi}\int_A \frac{(ikr - 1)(\mathbf{n} \times \mathfrak{E}) \times \mathbf{r}_1}{r^2}\, e^{ikr}\, d\Sigma, \qquad \ldots\ldots (3.21)$$

where $\mathbf{r}_1$ is the unit vector in the direction from the field point to the source point. Equation (3.21) expresses the electric field behind the aperture in terms of the assumed values of $\mathbf{E}_{\tan}$ in the aperture. For applications to slot radiators see Begovich (1950) and Severin (1950 a). For additional literature to the theory summarized above, see Bethe (1944), Caprioli (1949), Levine and Schwinger (1950), Luneberg (1944), Severin (1951), Silver (1949), Simon (1951) and Toraldo di Francia (1950 b).

It is almost impossible to give a complete survey of all possible approaches to the solution of electromagnetic diffraction problems. Pure integral equations have been given by Levine and Schwinger (1950) using the tensor free-space Green's functions (see also Severin 1951) but the kernels of the integral equation

are not always integrable. Other auxiliary wave potentials than the vector potentials may be used, for example the Debye potentials (Meixner 1948b, Flammer 1952). It must be remarked, however, that a one-component Hertzian vector is not sufficient. Bekefi's solution (Bekefi 1952a, b) has wrong singularities at the edge. In fact Bekefi's result is identical with the earlier one of Möglich (1927). It is claimed by Bekefi that the theoretically incorrect solution is still a good approximation to the correct solution for large apertures and for field points not too near the aperture.

In concluding this section we mention that the cross section theorem of Levine and Schwinger (eqn. (2.22)) can be extended to electromagnetic diffraction problems (Toraldo di Francia 1950a, Levine and Schwinger 1950).

§ 4. Kirchhoff's Scalar Theory of Diffraction and its Various Modifications

For a proper discussion of recent activity in the theory of diffraction it is necessary to recall first the basic part of Kirchhoff's theory. Let Σ be a screen of vanishing thickness covering a finite part of the plane $z=0$. Consider a system of sources in the left half-space ($z<0$). If the screen were absent these sources would produce a wave field $u_0(\mathrm{P})$, say, at the point P. The actual field $u(\mathrm{P})$ is the sum of $u_0(\mathrm{P})$ and $u_{\mathrm{diff}}(\mathrm{P})$, the diffracted field due to the secondary sources on Σ. By Green's theorem

$$u_{\mathrm{diff}}(\mathrm{P}) = \frac{1}{4\pi}\int\left\{u\frac{\partial}{\partial\nu}\left(\frac{e^{ikr}}{r}\right) - \frac{e^{ikr}}{r}\frac{\partial u}{\partial\nu}\right\}d\Sigma,$$

where the integration is over both faces of Σ and r is the distance from P to $d\Sigma$. Further, $\partial/\partial\nu$ denotes differentiation with respect to the integration-point coordinates in the direction of the normal to Σ drawn into free space. Unfortunately the secondary sources on Σ are not known. In accordance with Kirchhoff the following assumptions seem reasonable. First, on the illuminated face of Σ (in the sense of geometrical optics) the values of u and $\partial u/\partial\nu$ may be replaced by the corresponding unperturbed values, u_0 and $\partial u_0/\partial\nu$, as if the screen were absent. Secondly, on the dark face of Σ both u and $\partial u/\partial\nu$ may be set equal to zero. The total field then becomes, in Kirchhoff's approximation,

$$K_s(\mathrm{P}) = u_0(\mathrm{P}) + \frac{1}{4\pi}\int_S\left\{\frac{e^{ikr}}{r}\frac{\partial u_0}{\partial n} - u_0\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)\right\}d\Sigma, \quad \ldots\ldots(4.1)$$

where now the integration is over the illuminated part S of Σ only, while n refers to the normal of S drawn into the shadow region ($z>0$).

Serious objections can be raised against Kirchhoff's theory (Baker and Copson 1950, p. 71). In fact, eqn. (4.1) fails to reproduce the assumed values u_0 and $\partial u_0/\partial n$ when P tends to Σ. The limiting values of $K_s(\mathrm{P})$ are easily shown to be

$$Cu_0 + \frac{1}{4\pi}\int_S\frac{e^{ikr}}{r}\frac{\partial u_0}{\partial n}d\Sigma, \quad \ldots\ldots(4.2)$$

where $C=\frac{1}{2}$ or $\frac{3}{2}$ according as P lies on the dark or on the illuminated face of Σ. Consequently, only if

$$\tfrac{1}{2}u_0 + \frac{1}{4\pi}\int_S\frac{e^{ikr}}{r}\frac{\partial u_0}{\partial n}d\Sigma = 0 \qquad (\mathrm{P\ on\ S})$$

will the limiting values of K_s be identical with the assumed values u_0. However, it is at once evident (Bouwkamp 1953) that this condition cannot be fulfilled for

arbitrary S and u_0. This shows that Kirchhoff's procedure is not self-consistent. The same conclusion follows from a consideration of the limiting values of $\partial K_a/\partial n$, which are

$$C\frac{\partial u_0}{\partial n}-\left(k^2+\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\frac{1}{4\pi}\int_S u_0\frac{e^{ikr}}{r}\,d\Sigma. \qquad \ldots\ldots(4.3)$$

The reason for the inconsistency of Kirchhoff's theory is that u and $\partial u/\partial \nu$ cannot simultaneously be prescribed on Σ (the wave equation, $\Delta u+k^2u=0$, is elliptic). Discontinuities along the rim are of no account, however. Even infinities do not necessarily lead to troubles so long as they are absolutely integrable.

If Kirchhoff's boundary conditions on Σ were exact, then u and $\partial u/\partial n$ would jump by amounts u_0 and $\partial u_0/\partial n$ respectively across S. In fact these jumps *are* produced by K_a, as may at once be verified from the expressions (4.2) and (4.3) for the limiting values on S. This is in accordance with Kottler's interpretation of Kirchhoff's formula (4.1), in that K_a is the rigorous solution, not of a boundary-value problem, but of a saltus problem (Baker and Copson 1950, p. 98).

We shall now briefly discuss the complementary problem. Let A be a finite aperture cut in an infinite plane screen. Assuming the same primary field u_0 we apply eqn. (4.1) where S means the complement of A. To avoid the slight difficulty that S is no longer finite, it may be necessary to suppose that the imaginary part of k is positive. Now, the integral over S is equal to that over S+A minus that over A. If $z>0$, the integral over S+A equals $-u_0(P)$; if $z<0$, this integral is zero. Accordingly the total field behind the aperture, in Kirchhoff's sense, is

$$K_a(P)=-\frac{1}{4\pi}\int_A\left\{\frac{e^{ikr}}{r}\frac{\partial u_0}{\partial n}-u_0\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)\right\}d\Sigma, \qquad \ldots\ldots(4.4)$$

and that in front of the aperture is

$$K_a(P)=u_0(P)-\frac{1}{4\pi}\int_A\left\{\frac{e^{ikr}}{r}\frac{\partial u_0}{\partial n}-u_0\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)\right\}d\Sigma, \qquad \ldots\ldots(4.5)$$

n referring to the normal of the aperture drawn into the shadow region ($z>0$). The analytic continuation of the function (4.4) through the aperture is precisely the function (4.5), and vice versa. There are no discontinuities in the aperture, where we have

$$\left.\begin{aligned} K_a(P)&=\tfrac{1}{2}u_0(P)-\frac{1}{4\pi}\int_A\frac{e^{ikr}}{r}\frac{\partial u_0}{\partial n}\,d\Sigma,\\ \frac{\partial K_a(P)}{\partial n}&=\frac{1}{2}\frac{\partial u_0(P)}{\partial n}+\left(k^2+\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\frac{1}{4\pi}\int_A u_0\frac{e^{ikr}}{r}\,d\Sigma. \end{aligned}\right\} \qquad \ldots\ldots(4.6)$$

Equation (4.4) is usually derived in a different manner, namely, by applying Green's theorem for the half-space $z\geqslant 0$ and assuming that u and $\partial u/\partial n$ are zero on the dark face of the screen and equal to the unperturbed values *in the aperture*. However, Kirchhoff's original method should be preferred since it avoids the difficulty that (4.4) does not reproduce the values assumed in the aperture but rather the values (4.6).

For complementary problems (A = S) it follows from (4.1), (4.4) and (4.5) that $K_a+K_a=u_0$ ($z>0$) and $K_a+K_a=2u_0$ ($z<0$). This is Babinet's principle in the sense of Kirchhoff's theory (for details see Sommerfeld 1950, p. 208 ff.). The validity of Babinet's principle in Fraunhofer diffraction theory was doubted by Boersch (1951), but his argument is open to criticism.

It has often been suggested that Kirchhoff's formula gives the first term of an accurate solution of a boundary-value problem by successive approximations. This is wishful thinking, as was clearly shown by Franz (1949) and Schelkunoff (1951). In addition, Franz derived eqns. (4.4) and (4.5) as follows (a similar interpretation was given by Schelkunoff). A zero-order approximation is $u=u_0$ $(z<0)$, $u=0$ $(z>0)$. This agrees with the boundary condition at the (black) screen, but the wave equation is violated in the aperture. The wave equation is then restored by a correction term due to the secondary sources in the aperture. This term equals the right-hand side of eqn. (4.4), *for all* P $(z\geqslant 0)$. By elaborating the new interpretation, Franz (1949, 1950 a, b) devised an interesting diffraction theory which is applicable to all wavelengths. In the author's opinion, however, it is still an open question whether Franz's theory will eventually become of value in practice.

Various modifications of Kirchhoff's theory have recently been proposed for *planar* diffraction problems. The limitation to planar problems is essential since in the modified theory use is made of the two principal Green's functions of the half-space. Let $u(\mathrm{P})$ denote a wave function that is regular for $z\geqslant 0$. Then, in applying Rayleigh's formulae (2.6) and (2.7) to diffraction problems we shall assume, as in the usual version of Kirchhoff's theory, that u and $\partial u/\partial n$ may be replaced by the corresponding geometrical-optics values: immediately behind the screen u and $\partial u/\partial n$ are taken to be zero, and in the aperture they are replaced by u_0 and $\partial u_0/\partial n$. We have denoted Kirchhoff's solutions by K_a and K_s, where the subscripts refer to diffraction by a finite aperture and a finite screen respectively. It is convenient to introduce a similar notation for the wave functions derived from Rayleigh's formulae. For example, $R_{a1}(\mathrm{P})$ will denote the wave function for the diffraction by the *aperture* A based on Rayleigh's *first* formula (2.6). For obvious reasons, the corresponding modified Kirchhoff solutions will be termed 'Rayleigh solutions'. The complete set of Rayleigh solutions, for the diffracted field in the right half-space $(z>0)$, then becomes

$$R_{s1}(\mathrm{P})=u_0(\mathrm{P})+\frac{1}{2\pi}\int_S \frac{e^{ikr}}{r}\frac{\partial u_0}{\partial n}\,d\Sigma, \qquad \ldots\ldots(4.7)$$

$$R_{a1}(\mathrm{P})=-\frac{1}{2\pi}\int_A \frac{e^{ikr}}{r}\frac{\partial u_0}{\partial n}\,d\Sigma, \qquad \ldots\ldots(4.8)$$

$$R_{s2}(\mathrm{P})=u_0(\mathrm{P})-\frac{1}{2\pi}\int_S u_0\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)d\Sigma, \qquad \ldots\ldots(4.9)$$

$$R_{a2}(\mathrm{P})=\frac{1}{2\pi}\int_A u_0\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)d\Sigma. \qquad \ldots\ldots(4.10)$$

Unlike Kirchhoff's theory, the modified theory is self-consistent. The reason for this is that it was sufficient to assume boundary values for either u (case 2) or $\partial u/\partial n$ (case 1). In fact, all values assumed to exist at $z=+0$, whether behind the screen or in the aperture, are exactly reproduced by the Rayleigh solutions when the field point P tends to the plane $z=0$ from the right.

The analytic continuation of the Rayleigh solutions into the illuminated half-space is easy. Equations (4.7) and (4.9) remain valid for points P to the

left of the plane $z=0$. On the other hand eqns. (4.8) and (4.10) are to be replaced, in that order, by

$$R_{a1}(\mathrm{P}) = u_0(\mathrm{P}) - u_0(-\mathrm{P}) - \frac{1}{2\pi}\int_A \frac{e^{ikr}}{r}\frac{\partial u_0}{\partial n}\,d\Sigma, \qquad \ldots\ldots(4.11)$$

$$R_{a2}(\mathrm{P}) = u_0(\mathrm{P}) + u_0(-\mathrm{P}) + \frac{1}{2\pi}\int_A u_0\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)d\Sigma. \qquad \ldots\ldots(4.12)$$

Here $u_0(-\mathrm{P})$ means the value assumed by u_0 at the reflection of P in the plane $z=0$; $u_0(\mathrm{P}) \mp u_0(-\mathrm{P})$ is a zero-order reflected field in the sense of geometrical optics.

Like Kirchhoff's solution, the Rayleigh solutions are exact solutions of saltus problems. The functions R_{a2} and R_{a2} jump from $2u_0$ on the illuminated face of the screen to zero on the dark face. Similarly, the normal derivatives of R_{a1} and R_{a1} jump across the screen from $2\partial u_0/\partial n$ to zero. Further, the Kirchhoff solution is just the average of the two corresponding Rayleigh solutions, viz.

$$K_a = \tfrac{1}{2}(R_{a1} + R_{a2}), \qquad K_s = \tfrac{1}{2}(R_{a1} + R_{s2}),$$

while Babinet's principle now assumes either the form

$$R_{a1} + R_{a1} = u_0 \quad (z>0), \qquad R_{a1} + R_{a1} = 2u_0(\mathrm{P}) - u_0(-\mathrm{P}) \quad (z<0),$$

or the form

$$R_{a2} + R_{s2} = u_0 \quad (z>0), \qquad R_{a2} + R_{s2} = 2u_0(\mathrm{P}) + u_0(-\mathrm{P}) \quad (z<0).$$

One aim of the modified Kirchhoff theory was to provide a means for distinguishing between the two principal boundary-value problems, (I) $u=0$ on a *soft* screen and (II) $\partial u/\partial n = 0$ on a *rigid* screen, in the terminology of acoustics. Accordingly, since $\partial R_{a1}/\partial n = 0$ on the dark face of the screen, R_{a1} of eqn. (4.8) was proposed (Bouwkamp 1941) as an approximate expression for the diffracted field behind the aperture in the acoustically rigid screen for the incident wave u_0. Similarly, R_{a2} was suggested for the diffracted field behind the acoustically soft screen. It is obvious that these approximations will be accurate immediately behind the screen but poor in the vicinity of the aperture. The approximation is a complete failure if it is extended to the respective analytic continuations through the aperture into the illuminated space because on the lit face of the screen we have $\partial R_{a1}/\partial n = 2\partial u_0/\partial n$ and $R_{a2} = 2u_0$. In fact, the reflected-field terms in eqns. (4.11) and (4.12) suggest the opposite correspondence between the Rayleigh solutions and the solutions of the boundary-value problems.

An alternative way to show the close relation between the original and modified Kirchhoff solutions is the following. Consider, for example, the diffracted field behind an aperture. Noting that $\partial/\partial n = \partial/\partial z' = -\partial/\partial z$ for any function of r, we have from eqn. (4.4) that

$$2K_a(\mathrm{P}) = f(\mathrm{P}) + \frac{\partial}{\partial z}g(\mathrm{P}), \qquad \ldots\ldots(4.13)$$

where $f(\mathrm{P})$ and $g(\mathrm{P})$ are both even functions of z, viz.

$$f(\mathrm{P}) = -\frac{1}{2\pi}\int_A \frac{e^{ikr}}{r}\frac{\partial u_0}{\partial n}\,d\Sigma, \quad g(\mathrm{P}) = -\frac{1}{2\pi}\int_A \frac{e^{ikr}}{r}u_0\,d\Sigma. \qquad \ldots\ldots(4.14)$$

Comparison with (4.8) and (4.10) shows that

$$R_{a1}(\mathrm{P}) = f(\mathrm{P}), \qquad R_{a2}(\mathrm{P}) = \frac{\partial}{\partial z}g(\mathrm{P}). \qquad \ldots\ldots(4.15)$$

Consequently, the functions $R_{a1}(P)$ and $R_{a2}(P)$ are simply twice the even and odd parts respectively of $K_a(P)$. In addition, the function $f(P)$ is equal to the velocity potential of a membrane vibrating in a rigid baffle with velocity distribution $-\partial u_0/\partial n$; the same holds for $g(P)$ and $-u_0$. Methods and numerical results of the theory of acoustic radiation (Backhaus 1943, Bouwkamp 1946a, Carter and Williams 1951, 1952, Chetaev 1951a, b, Elrod 1952, Guptill 1953, Guptill and MacDonald 1952, King 1934, Molloy 1948, Nomura and Aida 1951, Nomura and Kawai 1949a, b, Pachner 1951, Schoch 1941, 1950, Stenzel 1942, 1949, 1952a, b) are therefore of value in diffraction theory also. Other authors that have used the modified Kirchhoff theory in one way or another are Bremmer (1951a, b), Di Jorio (1950a, b, 1951), Durand (1948a, b), Franz (1949, 1950a), Luneberg (1944), Scheffers (1942), Schoch (1950), Spence (1949), Swenson and Johnson (1952), and Toraldo di Francia (1949). Bremmer applied Rayleigh's second formula to the diffrac ion theory of gaussian optics. Various mathematical aspects of eqn. (2.7) were discussed by Luneberg, and Scheffers emphasized its usefulness, in the Fourier form, for the calculation of Fraunhofer patterns. Durand applied the Fourier equivalent of (4.10) to a circular aperture and to a half-plane for the case of a plane wave at normal incidence. Spence compared R_{a1} for a circular aperture (plane wave at normal incidence) with the corresponding exact solution of the boundary-value problem II. Experimental results on the diffraction of sound around a circular disc were discussed in connection with the Kirchhoff approximations by Primakoff, Klein, Keller and Carstensen (1947).

A slightly different modification of Kirchhoff's theory is as follows. If we assume that the unknown values of ϕ_1 and $\partial\phi_2/\partial n$ in the aperture (cf. eqns. (2.8) and (2.9)) may be replaced by the respective unperturbed values of the incident wave, we find $\phi_1 \simeq R_{a2}$, $\phi_2 \simeq R_{a1}$ $(z>0)$. If this is substituted in eqns. (2.1) and (2.2) we obtain

$$\left.\begin{aligned} u_{a1} &\simeq R_{a2} \quad (z>0), \qquad u_{a1} \simeq 2u_0 - R_{a2} \quad (z<0), \\ u_{a2} &\simeq R_{a1} \quad (z>0), \qquad u_{a2} \simeq 2u_0 - R_{a1} \quad (z<0). \end{aligned}\right\} \qquad \ldots\ldots(4.16)$$

For $z>0$ the approximation (4.16) is identical with that discussed in the beginning of this section. It should be noted that the approximate solutions of eqns. (4.16) do satisfy the correct boundary conditions at the screen. However, they are now not analytic functions: either the normal derivative or the approximate solution itself is discontinuous in the aperture.

In deriving the approximations (4.16), use has been made of the properties (ii) and (ii)′ of §2. Yet, almost equally characteristic for *planar* boundary-value problems are the properties (iv) and (iv)′. The latter express that for problem I ($u=0$ on the screen) $\partial u/\partial n$ is unperturbed in the aperture. Similarly, for problem II ($\partial u/\partial n=0$ on the screen) u is unperturbed in the aperture. Starting from this point of view we can derive a different set of approximations (Bekefi 1952c, 1953, Braunbek 1950a, Meixner and Fritze 1949, Severin and Starke 1952). From Rayleigh's formula (2.6) and property (iv) it follows that

$$\phi_1 = -\frac{1}{2\pi}\int_A \frac{e^{ikr}}{r}\frac{\partial u_0}{\partial n}\,d\Sigma - \frac{1}{2\pi}\int_S \frac{e^{ikr}}{r}\frac{\partial \phi_1}{\partial n}\,d\Sigma. \qquad \ldots\ldots(4.17)$$

Unfortunately, the values of $\partial\phi_1/\partial n$ are not known on the infinite screen S. If we assume that they are approximately zero, we arrive at $\phi_1 \simeq R_{a1}$. By a similar

reasoning in the case of problem II, we get $\phi_2 \simeq R_{a2}$. It is not difficult to verify that this ultimately gives

$$u_{a1} \simeq R_{a1}, \qquad u_{a2} \simeq R_{a2} \qquad \ldots\ldots(4.18)$$

everywhere in free space. The approximate solutions are now analytic and they produce the correct (unperturbed) values of $\partial u_{a1}/\partial n$ and u_{a2} in the aperture, but they violate the correct boundary conditions at the screen. In so far as an accurate approximation to the field is more important in the aperture than in the vicinity of the screen, the approximations (4.18) seem preferable to the opposite relations (4.16). On the other hand a linear combination of the Rayleigh solutions might give even better approximations (Meixner and Fritze 1949).

Approximate solutions for the diffraction by an aperture at once lead to approximate solutions for the diffraction by the complementary disc. The approximations corresponding to eqn. (4.16) are given by

$$u_{d1} \simeq R_{s1}, \qquad u_{d2} \simeq R_{s2} \qquad (z \gtrless 0), \qquad \ldots\ldots(4.19)$$

and the analogue of (4.18) is

$$\left.\begin{aligned} u_{d1} &\simeq R_{s2} \quad (z>0), \qquad u_{d1} \simeq 2u_0 - R_{s2} \quad (z<0) \\ u_{d2} &\simeq R_{s1} \quad (z>0), \qquad u_{d2} \simeq 2u_0 - R_{s1} \quad (z<0) \end{aligned}\right\} \qquad \ldots\ldots(4.20)$$

or, more explicitly,

$$u_{d1} \simeq u_0 \pm \frac{\partial}{\partial z}\frac{1}{2\pi}\int_S u_0 \frac{e^{ikr}}{r}\,d\Sigma \quad (z \gtrless 0), \qquad \ldots\ldots(4.21)$$

$$u_{d2} \simeq u_0 \pm \frac{1}{2\pi}\int_S \frac{\partial u_0}{\partial n}\frac{e^{ikr}}{r}\,d\Sigma \quad (z \gtrless 0). \qquad \ldots\ldots(4.22)$$

The last equation was applied by Sivian and O'Neil (1932), Muller, Black and Davis (1938), Wiener (1949 a), Meixner and Fritze (1949) for circular and square discs.

In addition to experimental verification, a common test object (e.g. Franz 1949) in diffraction theory is Sommerfeld's solution for the diffraction of a plane wave by an infinite half-plane. The exact solutions for a circular obstacle (of radius a), to be discussed later, may render the same service. There is no exact quantitative agreement between any of the Kirchhoff approximations and any of the exact solutions ($ka<10$). There is an indication that $u_{a1} \simeq R_{a1}$ is a better approximation than $u_{a1} \simeq R_{a2}$. Curves shown by Spence (1949) indicate that for $ka<5$ the approximation $u_{a2} \simeq R_{a1}$ is very poor in the aperture. The experimental work of Severin and Starke (1952) confirms that $u_{a2} \simeq R_{a2}$ (=their formula K) is much better. Detailed calculations show that the *main lobes* of the diffraction patterns for the saltus problems and the boundary-value problems are practically identical if $ka>5$, say. In spite of the extremely specialized problem of the circular aperture it is tentatively stated (Spence 1949) that this holds for any aperture provided the mean radius is nearly equal to or greater than the wavelength.

A great number of curves were given by Meixner and Fritze (1949) for the exact and Kirchhoff solutions of the diffraction of a plane wave (at normal incidence) by a rigid circular disc, or equivalently, for the radiation field of a freely vibrating circular plate. In this connection we recall (Bouwkamp 1941, Wiener 1951) that the velocity potential of the freely vibrating disc with velocity distribution $-\partial u_0/\partial x$ is given by $\mp\phi_1(x, y, \mp z)$, $z \lesseqgtr 0$, where ϕ_1 is the function defined before. Meixner and Fritze's function U is our u_{d2} for $u_0 = \exp(ikz)$,

and their approximations U_1 and U_2 (the less accurate) correspond to our equation (4.22) and the second part of (4.19) respectively, viz.

$$U_1 = e^{ikz} \pm \frac{ik}{2\pi}\int_S \frac{e^{ikr}}{r}\,d\Sigma, \qquad \ldots\ldots(4.23)$$

$$U_2 = e^{ikz} + \frac{1}{2\pi}\frac{\partial}{\partial z}\int_S \frac{e^{ikr}}{r}\,d\Sigma. \qquad \ldots\ldots(4.24)$$

In the sense of the modified Kirchhoff theory, the field of a freely vibrating circular disc was shown to be quite like that of the same disc vibrating in a rigid baffle. Further evidence of this, both theoretical and experimental, was provided by Wiener (1949 a, b, 1950) for circular and square plates. Experiments on diffraction by a circular aperture in a rigid screen were performed by Severin and Starke (1952), and the measured fields in the aperture and on the axis behind the aperture were compared with the Kirchhoff approximations for $0{\cdot}5 \leqslant a/\lambda \leqslant 4{\cdot}0$. As already mentioned, R_{a2} proved the better approximation to the rigorous solution.

In concluding this section, we shall briefly discuss a very interesting attempt by Braunbek (1950 a, b) to improve the modified Kirchhoff solution. Braunbek observed that the Kirchhoff solution does not constitute the main term (in a series of powers of $1/ka$) of the exact solution. To obtain the correct main term, we have to estimate the effect of the second integral of eqn. (4.17) on ϕ_1 (the problem for ϕ_2 is analogous). Braunbek replaced $\partial\phi_1/\partial n$ on the screen by the value following from Sommerfeld's theory for the half-plane, as if the screen S had locally a straight edge. This is a plausible assumption, because $\partial\phi_1/\partial n$ is expected to be rapidly decreasing from the edge over a distance of a few wavelengths (for an experimental confirmation in the case of ϕ_2 see Severin and Starke 1952). For the circular aperture and a normally incident plane wave Braunbek's approximation becomes

$$\phi_1 \simeq B_1 = -\frac{ik}{2\pi}\int_A \frac{e^{ikr}}{r}\,d\Sigma - \frac{ik}{2\pi}\int_S \frac{e^{ikr}}{r}[\Phi(k\xi) - \Psi(k\xi)]\,d\Sigma,$$

where ξ is the distance to the rim and

$$\Phi(x) = \frac{2}{\sqrt{\pi}}e^{-i\pi/4}\int_{\sqrt{x}}^{\infty} \exp(i\tau^2)\,d\tau, \qquad \Psi(x) = \frac{e^{i\pi/4}}{\sqrt{(\pi x)}}e^{ix}.$$

The evaluation of B_1 on the axis of the aperture is comparatively simple. Integration by parts and some trivial transformation yield

$$B_1(0,0,z) = e^{ikz} - \left(\frac{2}{\pi}\right)^{1/2}\exp(-i\pi/4)\exp\{ik(z^2+a^2)^{1/2}\} \times \int_0^{\infty}\left[\frac{\tau + ka + k(z^2+a^2)^{1/2}}{\tau + 2k(z^2+a^2)^{1/2}}\right]^{1/2}\frac{e^{i\tau}}{\sqrt{\tau}}\,d\tau. \qquad \ldots\ldots(4.25)$$

The integral is elementary if $z = 0$. At the centre of the aperture Braunbek's function is *exactly* equal to

$$B_1(0,0,0) = 1 - \sqrt{2}\,e^{ika}. \qquad \ldots\ldots(4.26)$$

When $z > 0$ the integral in (4.25) can be expanded asymptotically. To a first approximation, the first square root in the integrand may be replaced by its value at the lower limit of integration. Then, except for terms of order $1/ka$,

$$B_1(0,0,z) \sim e^{ikz} - \exp\{ik(z^2+a^2)^{1/2}\}\left[1 + \frac{a}{(z^2+a^2)^{1/2}}\right]^{1/2}, \qquad \ldots\ldots(4.27)$$

which is Braunbek's result obtained somewhat differently. His computations showed that (4.26) and (4.27) are in excellent agreement with Meixner and Fritze's (1949) exact values. The corresponding diffraction pattern was also evaluated by Braunbek.

In studying Braunbek's paper, the author encountered some difficulties in connection with a second type of approximation to ϕ_1. In starting from eqn. (2.8), Braunbek again replaces ϕ_1 in the aperture by the value $1-\Phi(k\xi)$ following from Sommerfeld's theory. Then

$$\phi_1 \simeq B_1{}^* = \frac{\partial}{\partial z}\left[-\frac{1}{2\pi}\int_A \{1-\Phi(k\xi)\}\frac{e^{ikr}}{r}\,d\Sigma\right].$$

Braunbek claims that, except for terms of order $1/ka$, the functions B_1 and $B_1{}^*$ are identical on the axis of the aperture. This, unfortunately, is incorrect, as may be seen from the respective values at the centre of the aperture. Without any calculations at all, it follows immediately that

$$B_1{}^*(0,0,0) = 1-\Phi(ka) \sim 1-\frac{e^{i\pi/4}}{\sqrt{(\pi ka)}}e^{ika},$$

which differs from (4.26).

§ 5. Approximations in Electromagnetic Diffraction Theory

Since rigorous solutions of electromagnetic diffraction problems can be obtained in a few simple cases only, it is of importance to have reasonable approximations at our disposal to deal with problems occurring in practice. Approximations have been known for a long time and were based on the work of Larmor, Tedone, Love, MacDonald, Bromwich, Von Ignatowsky, Tonolo, Kottler, and others (for references see Baker and Copson 1950, Vasseur 1952 a, b, c). The basic theorem underlying the approximations is some electromagnetic formulation of Huygens' principle or other, a so-called equivalence theorem (Schelkunoff 1936, 1939, 1951, Clavier 1948). One form of the equivalence theorem (Stratton and Chu 1939 a, b, Ledinegg 1948, Mirimanov 1950, Da Silveira 1951, de Broglie 1951, chap. VI) expresses the electromagnetic field in the interior of a source-free region inside a closed surface S (the sources of the field are assumed outside S) in terms of certain boundary values of the field at S, viz.

$$\left.\begin{aligned}\mathbf{E} &= \frac{1}{4\pi}\int_S \{(\mathbf{n}\times\mathbf{E})\times\nabla G + ik(\mathbf{n}\times\mathbf{H})G + (\mathbf{n}.\mathbf{E})\nabla G\}\,dS,\\ \mathbf{H} &= \frac{1}{4\pi}\int_S \{(\mathbf{n}\times\mathbf{H})\times\nabla G - ik(\mathbf{n}\times\mathbf{E})G + (\mathbf{n}.\mathbf{H})\nabla G\}\,dS,\end{aligned}\right\} \quad \ldots\ldots(5.1)$$

where $G = r^{-1}e^{ikr}$ while r denotes the distance between the field point (inside S) and the source point (at S). Further, $\mathbf{n}$ is the inward unit normal vector and ∇ is the gradient operator with respect to the integration-point coordinates. It should be noted that the integrals of eqns. (5.1) are identically zero for field points located outside S: there is no back radiation from the elementary sources.

Instead of G we might have used its conjugate $G^* = r^{-1}e^{-ikr}$ or any linear combination like $r^{-1}[\cos kr + A\sin kr]$, so long as S is finite. However, for

our choice of G eqns. (5.1) remain true in all possible cases. For example, if all sources of the electromagnetic field under consideration are *inside* a closed surface S, eqns. (5.1) hold for all points *outside* S on the understanding that $\mathbf{n}$ be the *outward* normal of S.

By purely mathematical transformation it can be shown (cf. Da Silveira 1951) that eqns. (5.1) are equivalent to

$$\mathbf{E} = \operatorname{curl}\mathbf{F} + ik\mathbf{A} - \operatorname{grad}\Psi, \qquad \mathbf{H} = \operatorname{curl}\mathbf{A} - ik\mathbf{F} - \operatorname{grad}\mathrm{X}, \quad \ldots\ldots(5.2)$$

where the auxiliary functions

$$\mathbf{A} = \frac{1}{4\pi}\int_S (\mathbf{n}\times\mathbf{H})G\,dS, \qquad \mathbf{F} = \frac{1}{4\pi}\int_S (\mathbf{n}\times\mathbf{E})G\,dS, \quad \ldots\ldots(5.3)$$

$$\Psi = \frac{1}{4\pi}\int_S (\mathbf{n}.\mathbf{E})G\,dS, \qquad \mathrm{X} = \frac{1}{4\pi}\int_S (\mathbf{n}.\mathbf{H})G\,dS, \quad \ldots\ldots(5.4)$$

may be interpreted physically† as electric vector potential $\mathbf{A}$, magnetic vector potential $\mathbf{F}$, electric scalar potential Ψ, and magnetic scalar potential X of, respectively, electric surface currents of density $\mathbf{I} = (c/4\pi)(\mathbf{n}\times\mathbf{H})$, magnetic surface currents of density $\mathbf{K} = -(c/4\pi)(\mathbf{n}\times\mathbf{E})$, electric surface charges of density $\sigma = (1/4\pi)(\mathbf{n}.\mathbf{E})$, and magnetic surface charges of density $\eta = (1/4\pi)(\mathbf{n}.\mathbf{H})$. Again, by pure mathematics, it follows that

$$\operatorname{div}\mathbf{A} = ik\Psi, \qquad \operatorname{div}\mathbf{F} = -ik\mathrm{X}, \qquad \ldots\ldots(5.5)$$

so that yet another equivalent form of eqns. (5.1) is obtained by elimination of the scalar potentials, viz.

$$\left.\begin{aligned}\mathbf{E} &= \operatorname{curl}\mathbf{F} - (1/ik)(k^2\mathbf{A} + \operatorname{grad}\operatorname{div}\mathbf{A})\\ &= \operatorname{curl}\mathbf{F} - (1/ik)\operatorname{curl}\operatorname{curl}\mathbf{A},\\ \mathbf{H} &= \operatorname{curl}\mathbf{A} + (1/ik)(k^2\mathbf{F} + \operatorname{grad}\operatorname{div}\mathbf{F})\\ &= \operatorname{curl}\mathbf{A} + (1/ik)\operatorname{curl}\operatorname{curl}\mathbf{F}.\end{aligned}\right\} \quad \ldots\ldots(5.6)$$

Let us now consider the diffraction of an arbitrary incident field $\mathbf{E}^i$, $\mathbf{H}^i$ (with sources in the left half-space $z<0$) by an aperture A in an infinite plane screen at $z=0$. In principle we could calculate the diffracted field behind the aperture ($z>0$) by applying any of the eqns. (5.1), (5.2) and (5.6), where S is taken as the plane $z=+0$ completed to a closed surface by the half-sphere at infinity at $z>0$. On account of the vector equivalent of Sommerfeld's radiation condition (e.g. Jones 1952e), the contribution of the half-sphere to the various integrals is zero, so that we are left with integrals over the aperture and the back face of the screen. By analogy with Kirchhoff's approximation in scalar diffraction theory, we assume that the actual field in the aperture (which is not known) may be replaced by the unperturbed incident field, and similarly that the field

† Equations (3.1) are extended as follows:

$$\operatorname{curl}\mathbf{E} = ik\mathbf{H} - (4\pi/c)\mathbf{K}, \qquad \operatorname{curl}\mathbf{H} = -ik\mathbf{E} + (4\pi/c)\mathbf{I},$$

where $\mathbf{K}$ and $\mathbf{I}$ are the magnetic and electric current densities respectively. The corresponding vector potentials are defined by the volume integrals

$$\mathbf{F} = -\frac{1}{c}\int \mathbf{K}G\,dV, \qquad \mathbf{A} = +\frac{1}{c}\int \mathbf{I}G\,dV.$$

Volume and surface current densities will not be distinguished in notation.

immediately behind the screen may be taken to be zero. On the basis of eqns. (5.3) and (5.6) we thus get the approximate solution

$$\left.\begin{aligned}\mathbf{E}_K^{\,d} &= \operatorname{curl}\frac{1}{4\pi}\int_A (\mathbf{n}\times\mathbf{E}^i)G\,dS - \operatorname{curl}\operatorname{curl}\frac{1}{4\pi ik}\int_A (\mathbf{n}\times\mathbf{H}^i)G\,dS,\\ \mathbf{H}_K^{\,d} &= \operatorname{curl}\frac{1}{4\pi}\int_A (\mathbf{n}\times\mathbf{H}^i)G\,dS + \operatorname{curl}\operatorname{curl}\frac{1}{4\pi ik}\int_A (\mathbf{n}\times\mathbf{E}^i)G\,dS,\end{aligned}\right\} \quad \ldots\ldots(5.7)$$

where $\mathbf{n}$ is the unit vector in the positive z direction while the field point is situated in the half-space $z>0$ behind the aperture. Since the integrals occurring in eqns. (5.7) are solutions of the vector wave equation for any given functions $\mathbf{E}^i$ and $\mathbf{H}^i$, it is obvious that the field $\mathbf{E}_K^{\,d}, \mathbf{H}_K^{\,d}$ defined by eqns. (5.7) does satisfy Maxwell's equations when $z>0$. However, the limiting values of $\mathbf{E}^d$ and $\mathbf{H}^d$ as z tends to zero from the right are not identical with the values on the plane $z=+0$ assumed in the derivation of eqns. (5.7) so that, as in the analogous scalar theory, the procedure leading to (5.7) is not self-consistent. Further, by analytic continuation of $\mathbf{E}_K^{\,d}, \mathbf{H}_K^{\,d}$ into the half-space $z<0$ through the aperture, the total field so obtained is in fact the rigorous solution of a saltus problem. Accordingly, the approximation (5.7) can never be considered as a first step toward an exact solution of a boundary-value problem (cf. Schelkunoff 1951)†.

Let us make similar approximations in eqns. (5.2), (5.3) and (5.4), that is, replace $\mathbf{E}$ and $\mathbf{H}$ in the integrals (5.3) and (5.4) by the value zero on the screen and by $\mathbf{E}^i$ and $\mathbf{H}^i$ in the aperture. A little surprisingly, it is then found that the field following from eqns. (5.2) is inconsistent with Maxwell's equations. The reason is that eqns. (5.5) are violated. Because the latter equations are equivalent to the continuity equations expressing the preservation of charge, additional terms in the scalar potentials due to line charges along the edge of the aperture become necessary. These line integrals can most easily be obtained by defining Ψ and X in terms of the approximations for $\mathbf{A}$ and $\mathbf{F}$ according to eqns. (5.5), viz.

$$\Psi_K = \frac{1}{4\pi ik}\operatorname{div}\int_A (\mathbf{n}\times\mathbf{H}^i)G\,dS, \qquad \mathrm{X}_K = -\frac{1}{4\pi ik}\operatorname{div}\int_A (\mathbf{n}\times\mathbf{E}^i)G\,dS.$$

By a simple mathematical transformation, and using the fact that $\mathbf{E}^i$, $\mathbf{H}^i$ satisfy Maxwell's equations, we next obtain

$$\left.\begin{aligned}\Psi_K &= \frac{1}{4\pi}\int_A (\mathbf{n}\,.\,\mathbf{E}^i)G\,dS + \frac{1}{4\pi ik}\int_C (\mathbf{s}\,.\,\mathbf{H}^i)G\,ds,\\ \mathrm{X}_K &= \frac{1}{4\pi}\int_A (\mathbf{n}\,.\,\mathbf{H}^i)G\,dS - \frac{1}{4\pi ik}\int_C (\mathbf{s}\,.\,\mathbf{E}^i)G\,ds,\end{aligned}\right\} \quad \ldots\ldots(5.8)$$

where $\mathbf{s}$ is the unit vector tangential to the edge C in the positive direction. The left-hand integrals present the contribution to the scalar potentials of the fictitious surface charges in the aperture, and the additional right-hand integrals are due to fictitious line charges of density $(1/4\pi ik)(\mathbf{s}\,.\,\mathbf{H}^i)$ and $(1/4\pi ik)(\mathbf{s}\,.\,\mathbf{E}^i)$, of electric and magnetic type respectively.

† "Auch das vektorielle Huygenssche Prinzip ist kein Zauberstab zur Lösung von Randwertproblemen", citation from Sommerfeld (1950, p. 336).

By substitution of (5.8) in (5.2) we find for our approximation

$$\left.\begin{aligned}\mathbf{E}_K{}^d &= \frac{1}{4\pi}\int_A \{(\mathbf{n}\times\mathbf{E}^i)\times\nabla G + ik(\mathbf{n}\times\mathbf{H}^i)G + (\mathbf{n}\,.\,\mathbf{E}^i)\nabla G\}\,dS \\ &\quad + \frac{1}{4\pi ik}\int_C (\mathbf{s}\,.\,\mathbf{H}^i)\nabla G\,ds, \\ \mathbf{H}_K{}^d &= \frac{1}{4\pi}\int_A \{(\mathbf{n}\times\mathbf{H}^i)\times\nabla G - ik(\mathbf{n}\times\mathbf{E}^i)G + (\mathbf{n}\,.\,\mathbf{H}^i)\nabla G\}\,dS \\ &\quad - \frac{1}{4\pi ik}\int_C (\mathbf{s}\,.\,\mathbf{E}^i)\nabla G\,ds,\end{aligned}\right\} \quad \ldots\ldots(5.9)$$

which should be compared with eqns. (5.1).

The necessity of introducing additional line integrals in the approximate Kirchhoff-like solution of electromagnetic diffraction by a (black) screen was first recognized by Kottler. Kottler's original equations, viz.

$$\left.\begin{aligned}\mathbf{E}_K{}^d &= \frac{1}{4\pi}\int_A \left(\mathbf{E}^i\frac{\partial G}{\partial n} - G\frac{\partial \mathbf{E}^i}{\partial n}\right)dS + \frac{1}{4\pi}\int_C (\mathbf{s}\times\mathbf{E}^i)G\,ds \\ &\quad + \frac{1}{4\pi ik}\int_C (\mathbf{s}\,.\,\mathbf{H}^i)\nabla G\,ds, \\ \mathbf{H}_K{}^d &= \frac{1}{4\pi}\int_A \left(\mathbf{H}^i\frac{\partial G}{\partial n} - G\frac{\partial \mathbf{H}^i}{\partial n}\right)dS + \frac{1}{4\pi}\int_C (\mathbf{s}\times\mathbf{H}^i)G\,ds \\ &\quad - \frac{1}{4\pi ik}\int_C (\mathbf{s}\,.\,\mathbf{E}^i)\nabla G\,ds,\end{aligned}\right\} \quad \ldots\ldots(5.10)$$

are fully equivalent to eqns. (5.7) or (5.9), and follow from them by suitable mathematical transformations. Kottler's equations (5.10) illustrate that omission of the line integrals (in other words, application of the scalar Kirchhoff procedure to the rectangular components of **E** and **H**) is not legitimate. The field defined by the line integrals alone does not satisfy Maxwell's equations, nor does the Kirchhoff part.

Many authors have recently devoted their attention to the classical Kirchhoff–Kottler formulation of the Huygens principle and its application to practical problems. Apart from the applications, essentially new knowledge has not been obtained. Old results were often re-formulated, re-interpreted, re-derived, re-named, or just simplified (?). Some authors find it advantageous to eliminate all physical considerations, while others do not like mathematical reasoning at all. Various applications of Kottler's formulae, including the radiation of wave guides and horns, and paraboloidal antennae, are scattered through the recent literature in the current technical periodicals. Many detailed calculations have been performed for small apertures of various shapes. Much useful material will be found, apart from the papers already cited, in those of Aoki (1949 b), Barrow and Chu (1939), Brodin (1949 a, b, 1950 a, b), Chu (1940), Croze and Boillet (1949), Croze and Darmois (1949), Croze and Durand (1949), De Possel and Pouget-Michel (1951), Durand (1948 d, 1949), Franz (1948, 1949, 1952), Goudet (1947, de Broglie 1948, p. 133), Gutton (1946), Horton (1949 b), Ortusi (1945), Ortusi and Simon (1950), Novobatzky (1942), Oseen (1944), Rocard (1946), Schouten and De Hoop (1952), Vasseur (1948, 1952 d), and Zuhrt (1947 a, b, 1948). Brodin's papers cited above are different from the others, in that Brodin is concerned with existence and uniqueness theorems in connection with Huygens' principle (reduction of the problem to integral equations).

Kottler's equations do not necessarily require that the aperture be cut in a *plane* screen. The modifications to be discussed presently are valid only for plane apertures and screens. The purpose of the modified theory is twofold: (*a*) to avoid inconsistencies in the derivation, and (*b*) to adapt the approximations to perfectly conducting screens. In the first place we observe from eqn. (3.18) that the field in the half-space $z>0$ can be expressed in terms of the tangential components of $\mathbf{H}$ at $z=+0$ alone. These components are known to be identical with the corresponding components of the unperturbed incident field. They are not known on the dark face of the screen. On the assumption that they are negligible at the screen, we find the approximation (Severin 1946, Ehrlich 1951, Vasseur 1952 d)

$$\left.\begin{aligned}\mathbf{E}_1^{\,d} &= \frac{1}{2\pi}\int_A \{ik(\mathbf{n}\times\mathbf{H}^i)G+(\mathbf{n}\,.\,\mathbf{E}^i)\nabla G\}\,dS+\frac{1}{2\pi ik}\int_C (\mathbf{s}\,.\,\mathbf{H}^i)\nabla G\,ds,\\ \mathbf{H}_1^{\,d} &= \frac{1}{2\pi}\int_A (\mathbf{n}\times\mathbf{H}^i)\times\nabla G\,dS.\end{aligned}\right\}\quad\ldots\ldots(5.11)$$

In the second place, we can express the diffracted field in terms of the tangential electric components, see eqn. (3.20). These components are zero on the screen but unknown in the aperture. Assuming that in the aperture they may be replaced by the tangential components of the incident electric field, we get (Severin 1951, Vasseur 1952 d)

$$\left.\begin{aligned}\mathbf{E}_2^{\,d} &= \frac{1}{2\pi}\int_A (\mathbf{n}\times\mathbf{E}^i)\times\nabla G\,dS,\\ \mathbf{H}_2^{\,d} &= \frac{1}{2\pi}\int_A \{-ik(\mathbf{n}\times\mathbf{E}^i)G+(\mathbf{n}\,.\,\mathbf{H}^i)\nabla G\}\,dS-\frac{1}{2\pi ik}\int_C (\mathbf{s}\,.\,\mathbf{E}^i)\nabla G\,ds.\end{aligned}\right\}\quad\ldots\ldots(5.12)$$

It should be noted that both approximations satisfy Maxwell's equations when $z>0$ and that their arithmetic mean is identical with Kottler's approximation for plane obstacles. The first approximation ($\mathbf{E}_1^{\,d}$, $\mathbf{H}_1^{\,d}$) does not satisfy the proper boundary conditions at the screen but it is in accordance with the tangential magnetic field being undisturbed in the aperture. On the other hand, the second approximation ($\mathbf{E}_2^{\,d}$, $\mathbf{H}_2^{\,d}$) has vanishing tangential electric field at the screen but the approximation seems to be poor in the aperture (for experimental evidence, see Andrews 1947). Of the three approximations discussed, only Kottler's is in agreement with Huygens' principle, which requires that there shall be no radiation in the backward direction of the wave front. Approximations for complementary discs follow from Babinet's principle (Severin 1950 b, 1951).

In view of eqns. (5.7) it is obvious that equivalent forms of eqns. (5.11) and (5.12) are given by

$$\left.\begin{aligned}\mathbf{E}_1^{\,d} &= -\frac{1}{2\pi ik}\,\mathrm{curl}\,\mathrm{curl}\int_A (\mathbf{n}\times\mathbf{H}^i)G\,dS,\\ \mathbf{H}_1^{\,d} &= \frac{1}{2\pi}\,\mathrm{curl}\int_A (\mathbf{n}\times\mathbf{H}^i)G\,dS,\end{aligned}\right\}\quad\ldots\ldots(5.13)$$

and

$$\left.\begin{aligned}\mathbf{E}_2^{\,d} &= \frac{1}{2\pi}\,\mathrm{curl}\int_A (\mathbf{n}\times\mathbf{E}^i)G\,dS,\\ \mathbf{H}_2^{\,d} &= \frac{1}{2\pi ik}\,\mathrm{curl}\,\mathrm{curl}\int_A (\mathbf{n}\times\mathbf{E}^i)G\,dS,\end{aligned}\right\}\quad\ldots\ldots(5.14)$$

respectively, so that the first approximation is based on an electric dipole layer and the second approximation on a magnetic dipole layer in the aperture. As a consequence, if eqns. (5.13) and (5.14) are applied to a plane polarized incident wave impinging on the aperture in the normal direction, the diffracted field is found to be given in terms of one scalar wave function and its derivatives. In the first case we have a one-component electric vector potential (or Hertzian vector) and in the second case a one-component magnetic vector potential (or Fitzgerald vector). More explicitly, for an incident wave of unit amplitude polarized parallel to the x axis we have

$$\mathbf{H}_1^{\,\mathrm{d}} = \operatorname{curl} \mathbf{A}, \qquad -ik\mathbf{E}_1^{\,\mathrm{d}} = \operatorname{curl}\operatorname{curl} \mathbf{A},$$

$$\mathbf{E}_2^{\,\mathrm{d}} = \operatorname{curl} \mathbf{F}, \qquad ik\mathbf{H}_2^{\,\mathrm{d}} = \operatorname{curl}\operatorname{curl} \mathbf{F},$$

where

$$-A_x = F_y = \frac{1}{2\pi}\int_{\mathrm{A}} \frac{\mathrm{e}^{ikr}}{r}\, dS, \qquad A_y = A_z = F_x = F_z = 0.$$

It therefore follows that the approximate solutions for the diffraction of a plane polarized wave normally incident on a circular aperture as given by Severin (1946), Ehrlich (1951), and Bekefi (1952 a, b) are all identical with that following from eqns. (5.11). These authors (see also Silver and Ehrlich 1951, Silver, Ehrlich and Held 1952) have found that the theoretical expressions derived for the diffracted field on the axis of the circular aperture are in close agreement with results of microwave measurements for apertures comparable with the wavelength. As in the rigorous solution, the diffracted field along the axis is polarized in the direction of the incident electric field:

$$H_{1y}^{\,\mathrm{d}} = \mathrm{e}^{ikz} - \frac{z}{\rho}\mathrm{e}^{ik\varrho},$$

$$E_{1x}^{\,\mathrm{d}} = \mathrm{e}^{ikz} - \frac{1}{2}\left(1 + \frac{z^2}{\rho^2} + \frac{a^2}{ik\rho^3}\right)\mathrm{e}^{ik\varrho},$$

where $\rho^2 = a^2 + z^2$ and a is the radius of the aperture. For a detailed discussion of this and the other approximations in relation to experiments the reader is referred to Vasseur (1952 d). Vasseur claims that the second approximation ($E_{2x}^{\,\mathrm{d}} = H_{1y}^{\,\mathrm{d}}$, $H_{2y}^{\,\mathrm{d}} = E_{1x}^{\,\mathrm{d}}$) is slightly better, contrary to Severin (1951). For earlier experiments, see Erler (1940) and Seiler (1931). Recent experiments on the diffraction pattern in and near a circular aperture have been reported by Andrews in a series of abstracts (Andrews 1946, 1948, 1950 a) and short papers (Andrews 1947, 1950 b, 1951 a). Related work is due to Hogg (1953), Houston and Noble (1951, 1952), Robinson (1953), and Tranter (1948). For diffraction patterns near rods, measurements have been reported and discussed by Andrews (1951 b) and Palmer (1952). Lack of space does not permit us to discuss in detail various other attempts to correlate diffraction measurements to theoretical approximations. In this connection we refer to the papers by Bekefi and Woonton (1952), Borts and Woonton (1952), Hartig (1950), Neugebauer (1952 a, b), Row (1952), Woodson (1942), Woonton (1950 a, b, 1952 a, b), and Woonton, Carruthers, Elliott and Rigby (1951). These papers contain a lot of experimental details. We finally want to mention two papers by Mirimanov (1948 c, 1949) on the diffraction of a spherical electromagnetic wave by a thin spherical segment and a circular disc, respectively, and the experimental study of Severin and Von Baeckmann (1951) on diffraction of microwaves by metallic and dielectric discs.

§ 6. Variational Formulation of Diffraction Problems

The development of variational techniques into a powerful tool for quantitative analysis of waveguide propagation and diffraction phenomena is largely due to Schwinger. His lectures (cf. Saxon 1945) at the Massachusetts Institute of Technology during the second world war constitute the basis of all that has been published since. Fortunately most of the results of Schwinger and his school are now readily available for practical application to electrical engineering (Marcuvitz 1951 a) and additions are planned for the near future by Levine and Schwinger in a monograph on the application of variational principles to diffraction problems.

A variational formulation of scalar diffraction problems for plane apertures, which permits approximate but accurate numerical evaluation of the diffracted amplitude and the transmission cross section for a wide range of frequencies, was given by Levine and Schwinger (1948 b, c). They illustrated the utility of the variational method by applying it to the circular aperture for a normally incident plane wave. The analysis was criticized by Copson (Baker and Copson 1950, p. 186) since many of the integrals involved appeared to diverge. Copson, however, in deriving what he calls "Levine and Schwinger's variational principle" in a mathematically sound way, confined himself to the problem for ϕ_2 (cf. § 2), while his criticism concerns that for ϕ_1. Fortunately the divergent integrals that occur in Levine and Schwinger's papers are easy to eliminate (Bouwkamp 1953) without affecting the numerical results.

Let $A_1(\mathbf{n}, \mathbf{n}')$ denote the amplitude of the diffracted wave ϕ_1 at large distances from the aperture, where $\mathbf{n}$ and $\mathbf{n}'$ are unit vectors in the direction of observation and of propagation of the incident plane wave respectively. Further, let θ and θ' be the angles between the positive z axis and these unit vectors. From eqn. (2.19) it follows that

$$A_1(\mathbf{n}, \mathbf{n}') = -(ik/2\pi)\cos\theta \int_A \phi_{n'}(\boldsymbol{\rho}') \exp[-ik\mathbf{n}\cdot\boldsymbol{\rho}']\, dS', \quad \ldots\ldots (6.1)$$

where $\phi_{n'}(\boldsymbol{\rho}')$ is the value of ϕ_1 in the aperture. Levine and Schwinger's integral equation contains a non-integrable kernel and should be interpreted in the sense of one of the differential-integral equations (2.11) or (2.12). Let us choose Maue's equation (2.12). Then

$$2\pi ik\cos\theta' \exp(ik\mathbf{n}'\cdot\boldsymbol{\rho}) = k^2 \int \phi_{n'}(\boldsymbol{\rho}') G(\boldsymbol{\rho}, \boldsymbol{\rho}')\, dS' - \int \nabla'\phi_{n'}(\boldsymbol{\rho}')\cdot\nabla' G(\boldsymbol{\rho}, \boldsymbol{\rho}')\, dS', \quad \ldots\ldots (6.2)$$

where $\boldsymbol{\rho}$ is in the aperture and $G(\boldsymbol{\rho}, \boldsymbol{\rho}') = |\boldsymbol{\rho} - \boldsymbol{\rho}'|^{-1} \exp[ik\,|\boldsymbol{\rho} - \boldsymbol{\rho}'|]$. If we multiply through in eqn. (6.2) by $\phi_{n''}(\boldsymbol{\rho})$ (where $\phi_{n''}(\boldsymbol{\rho})$ is the solution for a plane wave in the direction $\mathbf{n}''$), and integrate over the aperture, there results (after a slight transformation using the fact that $\phi_{n''}$ and $\phi_{n'}$ vanish at the edge):

$$2\pi ik\cos\theta' \int \phi_{n''}(\boldsymbol{\rho}) \exp(ik\mathbf{n}'\cdot\boldsymbol{\rho})\, dS = k^2 \int \phi_{n''}(\boldsymbol{\rho}) G(\boldsymbol{\rho}, \boldsymbol{\rho}') \phi_{n'}(\boldsymbol{\rho}')\, dS\, dS' - \int \nabla\phi_{n''}(\boldsymbol{\rho})\cdot\nabla'\phi_{n'}(\boldsymbol{\rho}') G(\boldsymbol{\rho}, \boldsymbol{\rho}')\, dS\, dS'. \quad \ldots\ldots (6.3)$$

The right-hand side of eqn. (6.3) is symmetrical in $\mathbf{n}'$ and $\mathbf{n}''$ and consequently so is the left-hand side. If we divide eqn. (6.3) by this left-hand side and by a

similar term in which $\mathbf{n}'$ and $\mathbf{n}''$ are interchanged, and then use eqn. (6.1) and invert, we obtain

$$A_1(\mathbf{n}'', \mathbf{n}') = A_1(-\mathbf{n}', -\mathbf{n}'')$$
$$= \frac{\cos\theta' \cos\theta'' \int \phi_{\mathbf{n}'}(\boldsymbol{\rho}) \exp(-ik\mathbf{n}''.\boldsymbol{\rho})\, dS \int \phi_{-\mathbf{n}''}(\boldsymbol{\rho}) \exp(ik\mathbf{n}'.\boldsymbol{\rho})\, dS}{\int \{\phi_{\mathbf{n}'}(\boldsymbol{\rho})\phi_{-\mathbf{n}'}(\boldsymbol{\rho}') - k^{-2}\nabla\phi_{\mathbf{n}'}(\boldsymbol{\rho}).\nabla'\phi_{-\mathbf{n}''}(\boldsymbol{\rho}')\} G(\boldsymbol{\rho}, \boldsymbol{\rho}')\, dS\, dS'} \quad \ldots\ldots(6.4)$$

It can be proved that the right-hand side of eqn. (6.4) is stationary with respect to small independent variations of $\phi_{\mathbf{n}'}$ and $\phi_{-\mathbf{n}''}$ about their *correct* values; only those variations which do not violate the condition $\phi_1 = 0$ at the edge are admissible.

The stationary character of the expression (6.4) is of importance for approximate calculations. In fact, a judicious choice of aperture distributions in eqn. (6.4) may result in a reasonably correct value for A_1 without the necessity of solving the original differential-integral equation. Scale factors are of no account since (6.4) is homogeneous of degree zero in ϕ. The same remarks apply to the transmission cross section which follows from eqn. (2.22):

$$\sigma_1(\mathbf{n}) = -\frac{2\pi}{k}\cos^2\theta\, \mathscr{I}\left[\frac{\int \phi_{\mathbf{n}}(\boldsymbol{\rho}) \exp(-ik\mathbf{n}.\boldsymbol{\rho})\, dS \int \phi_{-\mathbf{n}}(\boldsymbol{\rho}) \exp(ik\mathbf{n}.\boldsymbol{\rho})\, dS}{\int \{\phi_{\mathbf{n}}(\boldsymbol{\rho})\phi_{-\mathbf{n}}(\boldsymbol{\rho}') - k^{-2}\nabla\phi_{\mathbf{n}}(\boldsymbol{\rho}).\nabla'\phi_{-\mathbf{n}}(\boldsymbol{\rho}')\} G(\boldsymbol{\rho}, \boldsymbol{\rho}')\, dS\, dS'}\right] \quad \ldots\ldots(6.5)$$

Levine and Schwinger discussed the limiting form of this equation for low and for high frequencies. In the static case the leading term of the transmission cross section is given by

$$\sigma_1(\mathbf{n}) \sim \frac{2\pi}{3} k^4 \cos^2\theta \frac{\left[\int \phi(\boldsymbol{\rho})\, dS\right]^4}{\left[\int [\{\nabla\phi(\boldsymbol{\rho}).\nabla'\phi(\boldsymbol{\rho}')\}/\{|\boldsymbol{\rho} - \boldsymbol{\rho}'|\}]\, dS\, dS'\right]^2}, \quad (k \to 0) \quad \ldots\ldots(6.6)$$

which is an illustration of Rayleigh's λ^{-4} law for small scatterers.

The analogous problem for ϕ_2, discussed by Baker and Copson (1950, p. 186) and Levine (1950), is easier because it is based on the pure integral equation (2.10). From this equation it follows that

$$\int \psi_{\mathbf{n}'}(\boldsymbol{\rho}') G(\boldsymbol{\rho}, \boldsymbol{\rho}')\, dS' = -2\pi \exp(ik\mathbf{n}'.\boldsymbol{\rho}), \quad \ldots\ldots(6.7)$$

where $\psi_{\mathbf{n}'}(\boldsymbol{\rho}')$ is the value of $\partial\phi_2/\partial z$ in the aperture due to a plane wave incident in the direction $\mathbf{n}'$. The amplitude of the diffracted wave in the direction $\mathbf{n}$ is

$$A_2(\mathbf{n}, \mathbf{n}') = -\frac{1}{2\pi}\int \psi_{\mathbf{n}'}(\boldsymbol{\rho}') \exp(-ik\mathbf{n}.\boldsymbol{\rho}')\, dS', \quad \ldots\ldots(6.8)$$

and the analogue of eqn. (6.4) becomes

$$A_2(\mathbf{n}'', \mathbf{n}') = A_2(-\mathbf{n}', -\mathbf{n}'')$$
$$= \frac{\int \psi_{\mathbf{n}'}(\boldsymbol{\rho}') \exp(-ik\mathbf{n}''.\boldsymbol{\rho})\, dS \int \psi_{-\mathbf{n}''}(\boldsymbol{\rho}) \exp(ik\mathbf{n}'.\boldsymbol{\rho})\, dS}{\int \psi_{\mathbf{n}'}(\boldsymbol{\rho}) G(\boldsymbol{\rho}, \boldsymbol{\rho}') \psi_{-\mathbf{n}''}(\boldsymbol{\rho}')\, dS\, dS'}, \quad \ldots\ldots(6.9)$$

which expression, again, is stationary with respect to small variations of $\psi_{\mathbf{n}'}$ and $\psi_{-\mathbf{n}''}$ about their correct values. In this case the variations are not

restricted by a condition at the edge; the correct aperture fields ψ are infinite there.

The corresponding formula for the transmission cross section is

$$\sigma_s(\mathbf{n}) = -\frac{2\pi}{k}\mathscr{F}\left[\frac{\int \psi_n(\rho)\exp(-ik\mathbf{n}\cdot\rho)\,dS \int \psi_{-n}(\rho)\exp(ik\mathbf{n}\cdot\rho)\,dS}{\int\int \psi_n(\rho)G(\rho,\rho')\psi_{-n}(\rho')\,dS\,dS'}\right], \quad ..(6.10)$$

while the leading term in the static case is given by

$$\sigma_s(\mathbf{n}) \sim 2\pi \frac{\left[\int \psi(\rho)\,dS\right]^4}{\left[\int\int [\{\psi(\rho)\psi(\rho')\}/\{|\rho-\rho'|\}]\,dS\,dS'\right]^2}, \quad (k\to 0). \quad .....(6.11)$$

This leading term, it may be noted, is equal to $2\pi C^2$ where C is the electrostatic capacity of a metal disc of the same shape and size as the aperture.

The variational principle for ϕ_1 was successfully applied by Levine and Schwinger (1948 b) for the diffraction of a normally incident plane wave through a circular aperture by choosing as a trial aperture field

$$\phi_1 = a_1(1-\rho^2/a^2)^{1/2} + a_2(1-\rho^2/a^2)^{3/2}.$$

The values obtained in this way for the transmission cross section compare favourably with the exact values when $ka < 5$, say.

Alternatively, the variational principle may be applied to obtain a rigorous solution of the diffraction problem. The correct aperture distribution is assumed in the form (Sommerfeld 1942)

$$\phi_1 = \sum_{n=1}^{\infty} a_n(1-\rho^2/a^2)^{n-1/2}, \quad .....(6.12)$$

where the coefficients a_n have yet to be determined. Substitution of (6.12) in eqn. (6.4) gives (A_1 is the amplitude in the forward direction $\theta = 0$)

$$-\frac{A_1}{2a} = \frac{\left[\sum_{n=1}^{\infty}\{a_n/(2n+1)\}\right]^2}{\sum_{m=1}^{\infty}\sum_{n=1}^{\infty} c_{mn}a_m a_n}, \quad .....(6.13)$$

in which the coefficients c_{mn} are defined by

$$c_{mn} = (2/ka)^{m+n}\Gamma(m+\tfrac{1}{2})\Gamma(n+\tfrac{1}{2})f_{mn}(ka),$$

$$f_{mn}(\alpha) = \int_0^{\infty} (v^2-1)^{1/2} v^{-(m+n)} J_{m+1/2}(\alpha v) J_{n+1/2}(\alpha v)\,dv.$$

Let us apply small variations δa_n to the true values of a_n. In view of the stationary character of (6.13) we then find

$$\frac{1}{2m+1}\sum_{n=1}^{\infty}\frac{a_n}{2n+1} = -\frac{A_1}{2a}\sum_{n=1}^{\infty} c_{mn}a_n \qquad (m=1,2,\ldots).$$

On the other hand it follows from eqn. (6.1) that

$$A_1 = -ika^2\sum_{n=1}^{\infty}\frac{a_n}{2n+1}.$$

Elimination of A_1 thus gives the following infinite system of linear equation for the unknowns a_n:

$$\sum_{n=1}^{\infty} c_{mn}a_n = \frac{1}{ika(m+\frac{1}{2})} \qquad (m=1,2,\ldots). \quad .....(6.14)$$

This system was thoroughly studied by Magnus (1951, 1953). He showed that, for sufficiently small values of ka, the solution a_n is unique and can be expanded in a convergent power series in ka. Analytically simpler expansions and approximations for the cross section, by using Legendre functions, have been studied in detail elsewhere (Bouwkamp 1953) for both ϕ_1 and ϕ_2. A first-order approximation of the transmission coefficient in the case of ϕ_2 was calculated by Miles (1949 d) although this was done by a less powerful variational principle.

A second variational principle was derived by Levine and Schwinger (1949 a) for the diffracted amplitude $A_1(\mathbf{n}, \mathbf{n}')$ by considering the screen as an obstacle to the propagation of the incident wave through free space. This principle is based on the pure integral equation (2.14). Again, when applied to the case of the circular aperture for normal incidence, the numerical results obtained for the cross section closely agree with the exact values for $ka<5$. However, the leading term of the low-frequency cross section ($k\to 0$) was given by Levine and Schwinger to be

$$\sigma_1(\mathbf{n}) \sim \frac{k^2}{2\pi}\cos^2\theta \left[S_1 - \frac{\left(\int_{S_1}\chi(\boldsymbol{\rho})\,dS\right)\left(\int_{S_1}\chi(\boldsymbol{\rho})\,dS\int_{S_1}dS'/|\boldsymbol{\rho}-\boldsymbol{\rho}'|\right)}{\int_{S_1}\chi(\boldsymbol{\rho})\chi(\boldsymbol{\rho}')\,dS\,dS'/|\boldsymbol{\rho}-\boldsymbol{\rho}'|}\right]^2, \quad ..(6.15)$$

where S_1 is the area of the aperture S_1, and S_2 the infinite screen, while $\chi(\boldsymbol{\rho})$ is proportional to $\partial\phi_1/\partial z$. This leading term apparently departs from the k^4 behaviour characteristic of low frequencies, see eqn. (6.6). The solution, in our opinion, is that the right-hand side of eqn. (6.15), is identically zero for the correct low-frequency residual function χ, so that (6.15) seems to be of no value at all. In fact, it would be of interest to use for the circular aperture, instead of Levine and Schwinger's $\chi=(a/\rho)^3(1-a^2/\rho^2)^{-1/2}$, the residual function

$$\chi = -1+\frac{2}{\pi}\left[\arctan\left(\frac{\rho^2}{a^2}-1\right)^{1/2} - \left(\frac{\rho^2}{a^2}-1\right)^{-1/2}\right],$$

which corresponds to the 'static' field distribution on S_2. This function will also make the right-hand side of eqn. (6.15) vanish.

A similar second variational principle for the function ϕ_2 was discussed by Levine (1950). His analysis should be modified to eliminate divergent integrals.

Levine and Schwinger (1950) have extended their scalar variational techniques to electromagnetic diffraction by an aperture in an infinite plane conducting screen. Details need not be given here since the analysis is fully analogous to that for the scalar problems. Again, many of their equations have only a formal meaning because of the occurrence of divergent integrals. As before, the numerical results obtained for the plane-wave transmission cross section are not necessarily affected by this criticism.

To conclude this section, we emphasize once more the usefulness of variational techniques in diffraction problems. A multitude of technically important problems in waveguide and antenna engineering have been solved already. Without attempting to give a complete list of papers, we may mention in this connection the reports by Marcuvitz (1949, 1951 a) and Oliner (1951), as well as the papers by Kisun'ko (1949), Levine and Papas (1951), Miles (1946 a, 1948 c, 1949 c, e, 1950), Papas (1950), Sevick and Storer (1952), Sheingold (1953), Takahashi, Watanabe and Tanimoto (1950), and Tai (1952 a). In addition, Shmoys (1951) and Miles (1949 a) applied variational techniques to the diffraction of a plane wave through a grating, both for vanishing and

non-vanishing thickness of the grating elements. Sollfrey (1949) has given a general survey of the variational solution of scattering problems, and Levine (1951) calculated the resonant wavelength of a spherical cavity with a circular aperture. Finally, Lucke (1951b) has studied the transmission through a rectangular aperture by variational means, and Levitas and Lax (1951) the scattering and absorption by an acoustic strip.

§7. The Fourier Transform Method of Solving Integral Equations of Diffraction Theory

An amazingly powerful method of solving a general class of diffraction problems was recently developed by Schwinger and his school at the Massachusetts Institute of Technology. This method is an extension of the Wiener–Hopf technique (cf. Paley and Wiener 1934, Titchmarsh 1948) of the Fourier transform in the complex domain. Independently of Schwinger and collaborators, Copson (1946a) recognized the value of the new method by applying it to Sommerfeld's classical problem of diffraction by a half-plane. To give the reader an impression of how the new technique actually works, a short account of Copson's theory will be given. In fact it deals with the simplest of all problems that can be solved by the new technique.

Let the plane wave $u_0 = \exp[ik(y\cos\theta_0 + z\sin\theta_0)]$ impinge on the screen $y>0$, $z=0$, where θ_0 is the angle between the positive y axis and the direction of propagation of the incident wave; $0<\theta_0<\pi$. If $f(y)$ denotes the value of $(\partial/\partial z)(u_0-\phi_1)$ on the dark face of the screen, eqn. (2.15) becomes†

$$\int_0^\infty f(y')H_0^{(1)}[k|y-y'|]\,dy' = 2i\exp(iky\cos\theta_0),\quad y\geqslant 0. \quad \ldots\ldots(7.1)$$

Characteristic features of the integral eqn. (7.1) are (i) values of $f(y)$ assumed in $y<0$ are irrelevant; (ii) the left-hand side of (7.1) is unknown when $y<0$; (iii) the kernel is singular, and it is a difference kernel, that is, it depends only on $y-y'$; (iv) the kernel is of exponential type, that is, it decreases exponentially to zero if $|y-y'|$ tends to infinity, at least if k is assumed to be complex with $\mathscr{I}(k)>0$. In order to make eqn. (7.1) manageable by complex Fourier transforms, we *define* $f(y)$ to be identically zero when $y<0$, $g(y)$ to be zero when $y<0$, and equal to $2i\exp(iky\cos\theta_0)$ when $y\geqslant 0$, and finally $h(y)$ to be zero when $y\geqslant 0$ and equal to the left-hand side of (7.1) when $y<0$. The extended integral equation then becomes

$$\int_{-\infty}^{\infty} f(y')H_0^{(1)}(k|y-y'|)\,dy' = g(y)+h(y),\quad -\infty<y<\infty. \quad \ldots\ldots(7.2)$$

Now, let $F(w)$, $G(w)$, $H(w)$ and $L(w)$ denote the Fourier transforms of $f(y)$, $g(y)$, $h(y)$ and $H_0^{(1)}(k|y|)$, respectively. For example,

$$G(w) = \int_{-\infty}^{\infty} g(y)\exp(-iwy)\,dy = \int_0^\infty 2i\exp\{-iy(w-k\cos\theta_0)\}\,dy = \frac{2}{w-k\cos\theta_0},$$

valid when $\mathscr{I}(w)<\mathscr{I}(k\cos\theta_0)$. Similarly,

$$L(w) = \int_{-\infty}^{\infty} H_0^{(1)}(k|y|)\exp(-iwy)\,dy = \frac{2}{(k^2-w^2)^{1/2}},$$

† Exactly the same integral equation is obtained for the second boundary-value problem if $f(y)$ is identified with the aperture values of $\partial\phi_s/\partial z$, as is seen from eqn. (2.10). The equivalence of the two problems is a consequence of Babinet's principle. For a different approach to the solution of eqn. (7.1) see Magnus (1941a).

if $-\mathscr{I}(k)<\mathscr{I}(w)<\mathscr{I}(k)$, where the square root is equal to k if $w=0$. The Fourier transform of our kernel function is, therefore, one-valued and analytic in a certain strip parallel to the real axis of the complex w plane.

If $y\geqslant 0$, we have $f(y)=ik\sin\theta_0\exp(iky\cos\theta_0)-f_1(y)$, where $f_1(y)$ is the value of $\partial\phi_1/\partial x$ behind the screen. In view of Sommerfeld's radiation condition we may assume that $f_1(y)$ is bounded when $y\geqslant\epsilon>0$, and we further assume that $f_1(y)$ is absolutely integrable when $0\leqslant y\leqslant\epsilon$. It then follows that $F(w)$ is regular if $\mathscr{I}(w)$ is less than the smaller of 0 and $\mathscr{I}(k\cos\theta_0)$; in addition $F(w)$ is bounded when $|w|\to\infty$ in $\mathscr{I}(w)\leqslant\epsilon$. Except for a possible pole at $w=k\cos\theta_0$ (namely when $\theta_0>\pi/2$), $F(w)$ is regular in the lower half-plane.

With the aid of the asymptotic expansion of the Hankel function, it is easy to show that $|h(y)|$ is of the same order of magnitude as $\exp\{\mathscr{I}(ky)\}$ when $-\infty<y\leqslant-\epsilon<0$. If it is assumed that $|h(y)|$ is integrable if $-\epsilon\leqslant y\leqslant 0$, it follows that $H(w)$ is regular when $\mathscr{I}(w)>-\mathscr{I}(k)$, and bounded if $\mathscr{I}(w)\geqslant-\mathscr{I}(k)+\epsilon$.

The Fourier transform of the integral equation (7.2) follows from the convolution (Faltung) theorem, viz.

$$F(w)L(w)=G(w)+H(w), \qquad \ldots\ldots(7.3)$$

valid in the domain $-\mathscr{I}(k)<\mathscr{I}(w)<\min(0,\mathscr{I}(k\cos\theta_0))$ which is common to the respective domains of regularity of the four functions F, G, H and L. The *two* unknown functions F and H cannot, of course, be uniquely determined from the *one* relation (7.3). At this stage the analytical properties of $L(w)$ come into play; that is, the distribution of its singularities and zeros in the complex w plane and its asymptotic behaviour at infinity.

Equation (7.3) may be converted into

$$\frac{F(w)}{(k-w)^{1/2}}-\frac{(k+k\cos\theta_0)^{1/2}}{w-k\cos\theta_0}=\tfrac{1}{2}(k+w)^{1/2}H(w)+\frac{(k+w)^{1/2}-(k+k\cos\theta_0)^{1/2}}{w-k\cos\theta_0}. \qquad \ldots\ldots(7.4)$$

Now the left-hand side is regular in the half-plane $\mathscr{I}(w)<\min(0,\mathscr{I}(k\cos\theta_0))$ and the right-hand side is regular in the half-plane $\mathscr{I}(w)>-\mathscr{I}(k)$. These two half-planes overlap; therefore either side of eqn. (7.4) is the analytic continuation of the other side. Both define one and the same entire function $P(w)$, without singularities in the finite part of the w plane. It is easy to verify that $P(w)$ is of the order of $|w|^{1/2}$ if $|w|\to\infty$. Hence $P(w)$ is a polynomial in w of degree not greater than $\frac{1}{2}$; that is, $P(w)$ is a constant. This constant appears to be zero since the left side of (7.4) tends to zero in the appropriate lower half-plane. It thus follows that

$$F(w)=\frac{(k-w)^{1/2}(k+k\cos\theta_0)^{1/2}}{w-k\cos\theta_0}, \quad H(w)=\frac{-2}{(k+w)^{1/2}[(k+w)^{1/2}+(k+k\cos\theta_0)^{1/2}]}. \qquad \ldots\ldots(7.5)$$

Once $F(w)$, the Fourier transform of $f(y)$, is known we get by the inversion theorem

$$f(y)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{(k-w)^{1/2}(k+k\cos\theta_0)^{1/2}}{w-k\cos\theta_0}e^{ivw}\,dw, \qquad \ldots\ldots(7.6)$$

where the path of integration passes *below* the branch point $w=k$ and the pole $w=k\cos\theta_0$ of the integrand. By closing the path of integration in the lower

half-plane we see that $f(y)=0$ if $y<0$, as it should. By closing it in the upper half-plane we find, when $y>0$,

$$f(y)=ik\sin\theta_0\exp(iky\cos\theta_0)+\frac{1}{2\pi}\int_{\infty}^{(k+)}\frac{(k-w)^{1/2}(k+k\cos\theta_0)^{1/2}}{w-k\cos\theta_0}e^{iyw}\,dw,$$

where the first term arises from the pole. The remaining contour integral is expressible in terms of Fresnel's integral. The solution of the integral equation (7.1) is then found to be

$$f(y)=2ik\sin\theta_0\exp(iky\cos\theta_0)\int_0^{\sqrt{(2ky)}\sin(\theta_0/2)}\exp(i\tau^2)\,d\tau-\left(\frac{2k}{\pi}\right)^{1/2}\cos\frac{\theta_0}{2}\,e^{-i\pi/4}\frac{e^{iky}}{y^{1/2}},\qquad\ldots\ldots(7.7)$$

from which it is easy to determine the wave function ϕ_1.

In practical applications of the Wiener–Hopf technique, the really difficult problem is the factorization of the Fourier transform of the kernel of the integral equation. In the general case $L(w)$ is to be represented in the form $L(w)=L_1(w)L_2(w)$ where L_1 and L_2 are regular and non-vanishing in the appropriate upper and lower half-planes, respectively, and of algebraic growth as $|w|\to\infty$. In Copson's example we simply have $L_1=1/(k+w)^{1/2}$, $L_2=1/(k-w)^{1/2}$, but in more complicated applications infinite products enter into the analysis.

Schwinger discovered that the special case of electromagnetic boundary-value problems where a plane wave is incident upon a number of semi-infinite perfectly conducting parallel planes or cylinders can be formulated in terms of one or more Wiener–Hopf integral equations and therefore can be solved rigorously. An illustrative example of this was treated in detail by Carlson and Heins (1947) who considered an infinite set of staggered, equally spaced, semi-infinite metallic plates excited by a plane electromagnetic wave impinging on the system in a direction normal to the free edges with the electric vector parallel to the edges. Owing to the periodicity (equal spacing of plates) the problem reduces to *one* integral equation, which was solved on the assumption that in the space between the plates one and only one propagating mode occurs, or no such modes at all. Special attention was paid to the physically interesting reflection and transmission coefficients. In a second paper (Heins and Carlson 1947) the same authors discussed the analogous problem when the magnetic vector is parallel to the edges. The last problem is mathematically identical with the corresponding acoustic diffraction problem where the plates are understood to be perfectly rigid, and was treated as such by Miles (1948 a). This paper is almost an exact copy (including minor errors!) of that of Heins and Carlson, except that Miles' discussion is not restricted to the case of normal incidence (this extension is almost trivial for *scalar* diffraction).

The non-integrable kernel in the cited papers of Heins and Carlson and of Miles should be eliminated by placing $\partial^2/\partial y^2$ in front of the integral sign! This also applies to the paper by Heins (1950 b) concerning the case in which there are two reflected waves and appropriate restrictions on the wavelength, angle of stagger, and angle of incidence. As a check, it was shown that there exists a relation between the magnitudes of the new transmission and reflection coefficients which can be calculated by an extension of Poynting's theorem. Numerical results for non-staggered plates are to be found elsewhere (Marcuvitz 1951 a, p. 289).

The radiation and transmission properties of a waveguide consisting of a pair of semi-infinite parallel plates were studied by Heins (1948 a, b) for the case in which the electric vector is parallel to the edges and there is only one propagating mode in the guide. Let a denote the distance between the plates, θ the angle between the direction of incidence of the plane wave and the plates, and h the wave number of the propagating mode ($h^2 = k^2 - (\pi/a)^2$), then the electric field E_y inside the guide at large distances from the edge is asymptotic to $A \sin(\pi x/a) \exp(ihz)$, where the amplitude A (relative to that of the incident wave) is given by

$$A = 2\left[\frac{k\pi(1+\cos\theta)\cos(\frac{1}{2}ka\sin\theta)}{a^2(k^2\cos^2\theta - h^2)(k+h)}\right]^{1/2} \exp\{i\psi + \tfrac{1}{2}a(k\cos\theta - h)\}.$$

The phase constant is equal to

$$\psi = \theta_1 + \theta_2 + \arctan\left(\frac{k-h}{k+h}\right)^{1/2} + \tfrac{1}{2}ka\left(1 - \frac{\theta}{\pi}\right)\sin\theta + \frac{a(h - k\cos\theta)}{2\pi}\left[3 + \log\left(\frac{\pi}{\gamma ka}\right)\right]$$

where $\log\gamma$ is Euler's constant and θ_1 and θ_2 are defined by the infinite series

$$\theta_1 = -\sum_{n=1}^{\infty}\left[\arcsin\frac{ha}{\pi\{(2n+1)^2 - 1\}^{1/2}} - \frac{ha}{\pi(2n+1)}\right],$$

$$\theta_2 = \sum_{n=1}^{\infty}\left[\arcsin\frac{ka\cos\theta}{\{\pi^2(2n+1)^2 - (ka\sin\theta)^2\}^{1/2}} - \frac{ka\cos\theta}{\pi(2n+1)}\right].$$

Heins' second paper (1948 b) deals with the radiating properties of the same configuration when the guide is excited internally by the principal mode, which is partly reflected at the open end. His analysis again needs slight changes because of the occurrence of non-integrable kernels. An independent treatment of the problem was given by Vaĭnštein (1948 a) in a simpler way. In Russia the Wiener–Hopf technique was first considered by Fock (1942, 1944) apparently independently of the Anglo-Saxon scientists. Vaĭnštein calculated the reflection coefficients and the radiation patterns for five different excitations corresponding to the modes E_{00}, E_{01}, E_{02}, H_{01} (that of Heins), and H_{02}. He also compared the radiation patterns with those following from Kirchhoff's theory; there is in general a good agreement except for the back radiation and near the critical frequencies of the waveguide. In a second paper Vaĭnštein (1948 b) derived asymptotic expressions for the radiation patterns when a/λ is large. In every case they approximate more accurately to the correct pattern than does Kirchhoff's formula, especially in the backward direction. Equivalent-circuit parameters for the parallel-plate system are to be found in Marcuvitz (1951 a, p. 179 ff.).

A very interesting application of the new technique concerns the radiation of sound from an unflanged circular pipe (Levine and Schwinger 1948 a) excited internally by the dominant mode. With the end of the pipe chosen as a reference plane, the reflection coefficient is given by $R = -|R|\exp(2ikl)$, where

$$|R| = \exp\left[-\frac{2ka}{\pi}\int_0^{ka}\frac{\arctan\{-J_1(x)/Y_1(x)\}}{x(k^2a^2 - x^2)^{1/2}}\,dx\right],$$

and

$$\frac{\pi l}{a} = \int_0^{ka}\frac{\log[\pi J_1(x)[\{J_1(x)\}^2 + \{Y_1(x)\}^2]^{1/2}]}{x(k^2a^2 - x^2)^{1/2}} - \int_0^{\infty}\frac{\log[2I_1(x)K_1(x)]}{x(k^2a^2 + x^2)^{1/2}}\,dx,$$

in which a is the radius of the pipe and the Bessel-function notation is that of

Watson. If $ka \ll 1$, the reflection coefficient assumes the form $R = -\exp(2ikl_0)$ where

$$\frac{l_0}{a} = -\frac{1}{\pi}\int_0^\infty \log\left[2I_1(x)K_1(x)\right]dx/x^2 \simeq 0{\cdot}6133$$

is the exact value of the end correction (experimental values from 0·58 to 0·66). The angular distribution of the emitted radiation was also calculated for dominant-mode propagation ($0 < ka < 3{\cdot}832$). A type of reciprocity theorem allows the determination of the energy absorbed from an externally incident plane wave. The range is reduced to $0 < ka < 1{\cdot}841$ in the case of oblique incidence as also non-symmetric modes are excited at the mouth of the pipe. Numerical results on the radiation impedance were applied by Molloy (1950). Similar acoustic radiation and scattering problems associated with a semi-infinite hollow rigid cylinder of *elliptical* cross section have been solved by Blass (1951).

Independently of Levine and Schwinger, but slightly earlier, Vaĭnshteĭn (1947) published his study on the reflection of a sound wave at the open end of a circular tube. He found $l_0/a = 0{\cdot}613$ for the end correction in the static limit, well below the theoretical value obtained for the flanged pipe. Extensions to symmetrical and non-symmetrical electromagnetic waves appeared in a number of papers by the same author (Vaĭnshteĭn 1948 c, 1949, 1950 a, b, 1951 a, b). These papers are clearly written and well worth studying. An English translation of the important papers of Vaĭnshteĭn will soon be available (Shmoys 1953). Numerical results based on unpublished work of Schwinger and collaborators may be found in Marcuvitz (1951 a, p. 196 ff.). Reflection and transmission properties of an open-ended circular pipe for H_{11}-mode propagation have also been studied in an exhaustive report (in Japanese) by Iijima (1952). Radiation from a flanged circular pipe has been studied by Buchholz (1943 b) and Zernov (1951 b).

Several other problems bearing on diffraction theory in two dimensions were recently solved by the Wiener–Hopf technique. For instance, the coupling of two acoustical ducts was treated by Heins and Feshbach (1947) and by Heins (1951). Two other papers by Heins (1948 c, 1950 a) deal with water waves over a channel of finite depth with a dock and with a submerged plane barrier respectively.

Bailin (1951) applied the Fourier transform method to an infinite circular guide with a concentric semi-infinite circular bifurcating cylinder in order to study the effect of such a discontinuity upon plane longitudinal waves. His rigorous results show that an approximate solution due to Miles (1947 b) is very poor, see also Miles (1950). Jones (1952 d) considered the diffraction by two parallel plates of *finite* length and infinite width. His analysis leads to integral equations of the type encountered by Fox (1948, 1949, 1952, see also Baker and Copson, 1950, p. 180) in the diffraction of sound *pulses* by plates and slits. For the diffraction by a thick semi-infinite plate, see Jones (1953).

Two further papers deserve special attention in this connection. A simple approach, based on a representation of the scattered field as an angular spectrum of plane waves, was indicated by Clemmow (1951). This representation leads directly to a pair of dual integral equations, which replaces the single integral equation of Schwinger's method. The second paper is a report by Karp (1950) which emphasized the intimate connection that exists between the Wiener–Hopf technique and the method of separation of variables. Both authors represent the unknown fields as a superposition of plane waves travelling in all directions,

including complex ones (cf. §2). Inasmuch as the field is of greater importance than the currents induced in the plates, Clemmov's method is of advantage, but there still remains the difficult problem of factorizing the Fourier transform of the kernel functions. According to Jones (1952c) simplification can also be achieved by taking the Fourier transform before applying the boundary conditions. Another application of the Wiener–Hopf technique was made by Bazer and Karp (1952), to the problem of propagation of plane electromagnetic waves past a shore line; see also, in this connection, a recent paper by Clemmov (1953).

The propagation of sound waves in an open-ended channel consisting of two semi-infinite parallel plates was studied by Chester (1950a), but he was anticipated by Heins and by Vaĭnshteĭn already quoted. For propagation of sound pulses in the same system, see Chester (1950b, c).

For theoretical and experimental work on metal-plate dielectrics, bearing on the general subject of this section, the reader may consult the papers by Sakurai (1950b), Berz (1951), Lengyel (1951), Ruze and Young (1951), and Cheney and Watson (1951).

§8. Scalar Diffraction by Small Apertures and Narrow Slits

Equation (2.10) is the simplest of all integral equations considered so far. It was applied by Rayleigh (1897a) for apertures of dimensions very small compared with the wavelength. Rayleigh showed that the amplitude of the diffracted wave, for unit amplitude of the incident field in the aperture, is then approximately equal to the electrostatic capacity of a metal disc of the same size and shape as the aperture. Higher-order calculations for a small circular aperture (of radius a) were carried out but unpublished by the present author for the case of a normally incident wave, $u_0 = \exp(ikz)$. It was found that

$$\frac{\partial \phi_s(x,y,0)}{\partial z} = -\frac{2}{\pi}(a^2-\rho^2)^{-1/2}\left[1-2\epsilon-\frac{1}{2}\epsilon^2\{\eta^2-8\}\right.$$
$$\left.-\frac{1}{9}\epsilon^3\{3\eta^2-10\pi^2+72\}-\frac{1}{72}\epsilon^4\{\eta^4+4(\pi^2-12)\eta^2-8(\pi^2-12)^2\}+\ldots\right],$$

where $\epsilon = ika/\pi$, $\eta = \pi\rho/a$, $\rho^2 = x^2+y^2$. The corresponding amplitude $A_s(\theta)$ of the diffracted spherical wave at large distances from the aperture (defined by $\phi_s \sim A_s R^{-1}\exp(ikR)$ as $R\to\infty$) is given by

$$(2\pi/a)A_s(\theta) = 1+\left(\frac{1}{3}-\frac{4}{\pi^2}-\frac{1}{6}\sin^2\theta\right)(ka)^2$$
$$+\left\{\frac{16}{\pi^4}-\frac{20}{9\pi^2}+\frac{1}{15}+\left(\frac{2}{3\pi^2}-\frac{1}{15}\right)\sin^2\theta+\frac{1}{120}\sin^4\theta\right\}(ka)^4$$
$$-\frac{2}{\pi}ika\left\{1+\left(\frac{4}{9}-\frac{4}{\pi^2}-\frac{1}{6}\sin^2\theta\right)(ka)^2\right\}+O(k^5a^5),$$

where θ is the polar angle with respect to the symmetry axis of the aperture.

By evaluating

$$t_s = \frac{1}{\pi a^2}\int|A_s|^2\,d\Omega = \frac{2}{a^2}\int_0^{\pi/2}|A_s(\theta)|^2\sin\theta\,d\theta$$

we can calculate the transmission coefficient up to and including terms of order $(ka)^4$. Actually, however, the next term is also known, from the author's

(Bouwkamp 1941) calculations in spheroidal coordinates. The complete result is

$$t_2 = \frac{8}{\pi^3}\left[1+\left(\frac{4}{9}-\frac{4}{\pi^2}\right)(ka)^2+\left(\frac{71}{675}-\frac{8}{3\pi^2}+\frac{16}{\pi^4}\right)(ka)^4 + \left(\frac{568}{33075}-\frac{1936}{2025\pi^2}+\frac{128}{9\pi^4}-\frac{64}{\pi^6}\right)(ka)^6+\ldots\right]$$

$$= 0{\cdot}810\,569[1+0{\cdot}039\,160(ka)^2-0{\cdot}000\,7489(ka)^4-0{\cdot}000\,2602(ka)^6+\ldots]. \quad (8.1)$$

The slightly more complicated differential-integral equation (2.11) was also solved (Bouwkamp 1950 a) for a small circular aperture, again for normal incidence only. The tedious calculations were much simplified by a systematic use of certain complicated integrals involving Legendre functions (Bouwkamp 1949 b). Assuming in the aperture the power-series expansion (in the variable ika)

$$\phi_1(\rho) = \sum_{n=0}^{\infty}(ika)^n \sum_{m=0}^{n/2 \text{ or } (n-3)/2} B_{n,m} P_{2m+1}[(1-\rho^2/a^2)^{1/2}]$$

we were able to find the numerical coefficients $B_{n,m}$ by a rather complicated recursive relation. The calculation was performed explicitly up to and including terms of order $(ka)^6$. In the equivalent form of eqn. (6.12) the coefficients a_n are found to be

$$a_1 = -\frac{2ika}{\pi}\left[1+\frac{1}{6}(ka)^2+\frac{2i}{9\pi}(ka)^3+\frac{1}{120}(ka)^4+\frac{13i}{225\pi}(ka)^5 + \left(\frac{1}{5040}-\frac{4}{81\pi^2}\right)(ka)^6+\frac{323i}{44100\pi}(ka)^7+\ldots\right],$$

$$a_2 = -\frac{i(ka)^3}{9\pi}\left[1+\frac{3}{10}(ka)^2+\frac{2i}{5\pi}(ka)^3+\frac{1}{56}(ka)^4+\frac{19i}{175\pi}(ka)^5+\ldots\right],$$

$$a_3 = -\frac{i(ka)^5}{300\pi}\left[1+\frac{5}{14}(ka)^2+\frac{10i}{21\pi}(ka)^3+\ldots\right],$$

$$a_4 = -\frac{i(ka)^7}{17640\pi}[1+\ldots].$$

The corresponding less complete expressions of Sommerfeld (1942, 1948 a) and Magnus (1951) contain errors (Bouwkamp 1953).

The relevant amplitude $A_1(\theta)$ of the diffracted wave at large distances from the aperture is known (Bouwkamp 1950 a) up to and including terms of order $(ka)^6$. Let it suffice, however, to quote the corresponding formula for the transmission coefficient, viz.

$$t_1 = \frac{8}{27\pi^2}(ka)^4\left[1+\frac{8}{25}(ka)^2+\frac{311}{6125}(ka)^4+\left(\frac{2612}{496125}-\frac{4}{81\pi^2}\right)(ka)^6+\ldots\right]$$

$$= 0{\cdot}030\,0211(ka)^4[1+0{\cdot}32(ka)^2+0{\cdot}050\,7755(ka)^4+0{\cdot}000\,2613(ka)^6+\ldots]. \quad \ldots\ldots (8.2)$$

The first term of this expansion has been known since Rayleigh. It was re-derived by King (1935) and Braumann (1948). The result of Hanson (1933) was in error (Gutin 1937). Equation (8.2) was obtained (Bouwkamp 1949 a) by three different methods, including the variational method of Levine and Schwinger (1949 b). The author is convinced that eqn. (8.1) for t_2 is error-free also, in spite of the laborious numerical calculations by which it was obtained.

A similar analysis might be carried out for an obliquely incident wave (cf. Miles 1952a, d), and a solution for elliptic apertures is not impossible but nothing beyond Rayleigh's first-order solution seems to have been accomplished thus far. It is also well to recall that these numerical results (adequate for applications if $ka<1$) are of no value in electromagnetic diffraction theory. The following is quoted (but for notation) from Rayleigh (1897b, p. 326): "It had at first been supposed that the solution ϕ_1 might be applied directly to E_x, which satisfies the condition of vanishing upon the faces of the screen. If this were admitted, as also $E_y=0$ throughout (?), the value of E_z would follow by $\operatorname{div}\mathbf{E}=0$. The argument was, however, felt to be sufficient on account of the discontinuities which occur at the edge of the aperture . . .".

For very narrow slits, $kb\ll 1$, first-order approximate solutions of eqns. (2.25) and (2.26) were derived by Rayleigh (1897a). In a later paper (Rayleigh 1913) he extended his calculations to slits slightly narrower than one wavelength, for a plane wave at normal incidence. For example, in the case of eqn. (2.25) Rayleigh assumed the aperture distribution

$$\frac{\partial\phi_2(y,0)}{\partial z}=\frac{P}{(1-y^2/b^2)^{1/2}}+Q,\qquad -b\leqslant y\leqslant b$$

in which P and Q are independent of y. On substituting this in the left-hand side of eqn. (2.25), Rayleigh evaluated the resulting complicated integrals up to and including terms of order $(kb)^6$ and then determined P and Q (as functions of $kb\leqslant 2$) such as to get a practically constant value [viz. $2iu_0(0,0)$] for the left-hand side of eqn. (2.25). It is obvious that this numerical solution is not a consistent analytical approximation to successive powers of kb.

The first two terms of a 'power-series' solution of eqn. (2.25) were given by Sommerfeld (1950, p. 294). The author checked Sommerfeld's solution and in addition calculated the third-order term (Bouwkamp 1953). On the assumption that $u_0=\exp(ikz)$, the result is

$$\frac{\partial\phi_2(y,0)}{\partial z}=\frac{1}{b}\sum_{n=0}^{\infty}C_n(1-y^2/b^2)^{n-1/2},$$

where

$$C_0=\frac{1}{p}-\frac{1}{4}\left(1-\frac{1}{2p}\right)(kb)^2+\frac{1}{128}\left(1-\frac{9}{4p}+\frac{1}{p^2}\right)(kb)^4+\ldots,$$

$$C_1=\frac{1}{2}\left(1+\frac{1}{2p}\right)(kb)^2-\frac{1}{16}\left(1-\frac{3}{4p}\right)(kb)^4+\ldots,$$

$$C_2=\frac{1}{48}\left(1+\frac{3}{4p}\right)(kb)^4+\ldots,$$

while the parameter p is defined by

$$p=\log(\gamma kb/4)-\pi i/2,\qquad \log\gamma=0{\cdot}57721\ldots\ \text{(Euler's constant)}.$$

The parameter p occurs because of the logarithmic singularity of the Hankel function. The above series are therefore no power series in the strict sense. This raises a question (Groschwitz and Hönl 1952) whether the expansions are convergent or merely asymptotic as $kb\to 0$. The answer to the analogous question for a circular aperture is in favour of the first alternative (Magnus 1951, 1953).

At large distances ρ from the centre of the slit we have

$$\phi_2 \sim \frac{1}{2p}\left(\frac{2\pi}{k\rho}\right)^{1/2} \exp\{i(k\rho - 3\pi/4)\}\left[1 + \frac{1}{4}(kb)^2\cos^2\theta + \frac{1}{64}(kb)^4\left\{-p\cos 2\theta + \sin^4\theta - 3\sin^2\theta + \frac{3}{4} + \frac{1}{2p}\right\} + \ldots\right],$$

where θ is the polar angle with respect to the x axis. The transmission coefficient of the slit appears to be

$$\tau_2 = \frac{\pi^2/kb}{\pi^2 + [2\log(\gamma kb/4)]^2}\left[1 + \frac{1}{4}(kb)^2 + \frac{3}{256}(kb)^4 \times \left\{1 + \frac{16}{3}\frac{\log(\gamma kb/4)}{\pi^2 + [2\log(\gamma kb/4)]^2}\right\} + \ldots\right] \quad \ldots\ldots(8.3)$$

the first two terms of which were given by Sommerfeld. Equation (8.3) gives also the plane-wave transmission coefficient of a narrow slit in a conducting screen for the case of normal incidence when the *magnetic* vector of the incident wave is parallel to the edges of the slit (x component of magnetic vector may be identified with ϕ_2).

Sommerfeld also discussed the diffraction by a narrow slit when the *electric* vector is parallel to the edges of the slit. In this case the x component of the electric vector may be represented by ϕ_1. For normal incidence the differential-integral equation (2.26) reduces to

$$\left(k^2 + \frac{d^2}{dy^2}\right)\int_{-b}^{b} \phi_1(y', 0) H_0^{(1)}(k|y - y'|)\, dy' = 2k, \quad \ldots\ldots(8.4)$$

which in the sense of eqn. (2.13) can be transformed into

$$\int_{-b}^{b} \phi_1(y', 0)[H_0^{(1)}(k|y - y'|) - \cos ky H_0^{(1)}(k|y'|)]\, dy' = \frac{2}{k}(1 - \cos ky). \quad \ldots\ldots(8.5)$$

By assuming a solution of eqn. (8.5) in the form

$$\phi_1(y, 0) = -ikb \sum_{n=1}^{\infty} D_n(1 - y^2/b^2)^{n-1/2},$$

Sommerfeld (1950, p. 295, eqn. (31)) found $D_1 = 1 - \frac{1}{4}(p - 3/2)(kb)^2 + \ldots$, $D_2 = -\frac{1}{12}(kb)^2 + \ldots$, which differs from the result more recently obtained by Groschwitz and Hönl (1952). Both results are in error (see also Hönl and Zimmer 1953). The correct expressions are

$$D_1 = 1 - \frac{1}{4}\left(p - \frac{1}{2}\right)(kb)^2 + \frac{1}{16}\left(p^2 - \frac{7}{8}p + \frac{7}{32}\right)(kb)^4 + \ldots,$$

$$D_2 = \frac{1}{12}(kb)^2 - \frac{1}{32}\left(p - \frac{1}{2}\right)(kb)^4 + \ldots,$$

$$D_3 = \frac{1}{320}(kb)^4 + \ldots.$$

At large distances from the slit the corresponding diffraction pattern is given by

$$\phi_1 \sim \frac{1}{4}(kb)^2\cos\theta\left(\frac{2\pi}{k\rho}\right)^{1/2} \exp\{i(k\rho - 3\pi/4)\}\left[1 - \frac{1}{4}\left\{p - \frac{3}{4} + \frac{1}{2}\sin^2\theta\right\}(kb)^2 + \frac{1}{32}\left\{2p^2 - \frac{5}{2}p + \frac{7}{8} + \left(p - \frac{2}{3}\right)\sin^2\theta + \frac{1}{6}\sin^4\theta\right\}(kb)^4 + \ldots\right],$$

and for the transmission coefficient (electric vector parallel to the edges) we have

$$\tau_1 = \frac{\pi^3}{32}(kb)^3\left[1+\frac{5}{16}(kb)^2\left\{1-\frac{8}{5}\log\left(\frac{\gamma kb}{4}\right)\right\}+\ldots\right]. \quad \ldots\ldots(8.6)$$

Equations (8.3) and (8.6), in their dependence on kb, confirm the experimental fact that, if natural light is incident on a very narrow slit, the diffracted field behind the aperture is substantially polarized perpendicular to the edge (for wider slits, $kb \simeq 2$, this is not true, cf. Rayleigh 1913). The scattering coefficients for a narrow *strip* of width $2b$ are now also known in virtue of Babinet's principle. They are equal to $2\tau_1$ and $2\tau_2$ if the incident light is polarized perpendicular and parallel, respectively, to the edges of the strip.

Sommerfeld's procedure, starting from eqn. (8.5), is simpler than that of Groschwitz and Hönl. Still simpler is to start with eqn. (8.4) and using the Fourier expansion of the static kernel $\log(2|\sin\theta-\sin\theta'|)$ (cf. Marcuvitz 1951 a, p. 147, Jaeckel 1951) or its equivalent expansion in Tschebysheff polynomials of argument $(1-y^2/b^2)^{1/2}$ where $y=b\sin\theta$. The analytic similarity with the problems for the circular aperture is then most obvious (Bouwkamp 1953). This also follows from the recent paper by Müller and Westpfahl (1953). It is gratifying to mention that these authors have arrived independently at the expressions (8.3) and (8.6) for the transmission coefficients.

In concluding this section we may refer to a paper by Ingerslev and Nielsen (1944) who extended Rayleigh's theory of small apertures so as to include the finite thickness of the wall. Diffraction phenomena due to a large assembly of circular apertures or discs have been studied by Reesinck and De Vries (1940), Kathavate (1945 b), Toraldo di Francia (1948), and lately by Grosjean (1952).

§ 9. Diffraction of a Plane Electromagnetic Wave by a Small Circular Obstacle

Bethe (1944) has given a detailed analysis of the diffraction of electromagnetic radiation by a circular hole small compared with the wavelength and he has emphasized that the result is completely different from that obtained by the Kirchhoff method (cf. Pekeris 1944). Later Copson (1946 b) re-derived Bethe's expressions for the field in the aperture and also obtained a solution for the case of the disc. Unfortunately the theories of Bethe and Copson are open to criticism (Bouwkamp 1950 e, 1953, Copson 1951) in that they do not predict the correct approximations for the field in the vicinity of the obstacle. Their expressions for the field at large distances from the obstacle, however, are correct. In addition, the present author has checked that Bethe's application of his theory to the coupling of cavities by a small circular hole is not necessarily affected by this criticism.

With Bethe we observe that the diffracted field behind the circular aperture in a plane conducting screen can be derived from fictitious magnetic currents and charges in the aperture. However, it is most convenient to eliminate from the discussion the scalar magnetic potential and to base the analysis upon eqns. (3.10), expressing the diffracted field in terms of the magnetic vector potential alone. This procedure will automatically render Bourgin's criticism on Bethe's theory (Bourgin 1945) irrelevant. The magnetic current density $\mathbf{K}$

must then be determined from a system of differential-integral equations that is the dual of the system (3.11), viz.

$$F_x(x,y,0) = -\frac{1}{c}\int K_x(x',y')\frac{e^{ikr}}{r}\,dx'\,dy', \qquad \ldots\ldots(9.1)$$

$$F_y(x,y,0) = -\frac{1}{c}\int K_y(x',y')\frac{e^{ikr}}{r}\,dx'\,dy', \qquad \ldots\ldots(9.2)$$

$$\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} = E_z^{\mathrm{i}}(x,y,0), \qquad \ldots\ldots(9.3)$$

$$(k^2+\partial^2/\partial x^2+\partial^2/\partial y^2)F_x = -\left[\frac{\partial}{\partial z}E_y^{\mathrm{i}}(x,y,z)\right]_{z=0}, \qquad \ldots\ldots(9.4)$$

$$(k^2+\partial^2/\partial x^2+\partial^2/\partial y^2)F_y = \left[\frac{\partial}{\partial z}E_x^{\mathrm{i}}(x,y,z)\right]_{z=0}, \qquad \ldots\ldots(9.5)$$

where r is the distance between the field point $(x,y,0)$ and the source point $(x',y',0)$, both located in the aperture.

Since a rigorous solution of this complicated system of differential-integral equations seems impossible, the term 'approximate solution' should be handled with care. In our opinion, the only feasible definition of an approximate solution is one that is based on a consistent power-series solution in the variable ka (where a is the radius of the aperture). It is not *a priori* obvious that replacing E_x^{i}, $\partial E_y^{\mathrm{i}}/\partial z$, $\partial E_z^{\mathrm{i}}/\partial z$ by their respective values at the centre of the aperture and replacing $r^{-1}\exp(ikr)$ by r^{-1}, as in the methods of Bethe and of Copson, will lead to a significant approximation of $\mathbf{K}$ correct up to and including terms of relative order ka, not to mention that the 'approximate' system of differential-integral equations may have more than one solution. In a self-consistent approximation we must expand the right-hand sides of eqns. (9.3), (9.4) and (9.5) in powers of ka and retain as many terms as required for the ultimate accuracy aimed at.

For example, let the incident field be a plane wave of unit amplitude with components

$$\mathbf{E}^{\mathrm{i}} = (\cos\phi_0\cos\theta_0,\ \sin\phi_0,\ -\cos\phi_0\sin\theta_0)\exp[ik(x\sin\theta_0 + z\cos\theta_0)],$$

$$\mathbf{H}^{\mathrm{i}} = (-\sin\phi_0\cos\theta_0,\ \cos\phi_0,\ \sin\phi_0\sin\theta_0)\exp[ik(x\sin\theta_0 + z\cos\theta_0)],$$

where the xz plane is chosen as the plane of incidence while ϕ_0 is the angle between $\mathbf{E}^{\mathrm{i}}$ and the xz plane and θ_0 the angle between the z axis and the direction of incidence. Equations (9.3) to (9.5) then become

$$\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} = -\cos\phi_0\sin\theta_0\exp(ikx\sin\theta_0), \qquad \ldots\ldots(9.6)$$

$$(k^2+\Delta)F_x = -ik\sin\phi_0\cos\theta_0\exp(ikx\sin\theta_0), \qquad \ldots\ldots(9.7)$$

$$(k^2+\Delta)F_y = ik\cos\phi_0\cos^2\theta_0\exp(ikx\sin\theta_0). \qquad \ldots\ldots(9.8)$$

Further, let $\mathbf{K}^0$ and $\mathbf{K}^1$ denote the terms of $\mathbf{K}$ of relative order zero and one in ka (or equivalently in k) respectively. It appears that in the case of plane-wave excitation relative order is the same as absolute order. Hence assume $\mathbf{K} = \mathbf{K}^0 + \mathbf{K}^1 + O(k^2)$, where $\mathbf{K}^0 = O(1)$, $\mathbf{K}^1 = O(k)$. Then, expanding the exponential function in eqns. (9.1) and (9.2) in powers of k, we get

$$\mathbf{F}(x,y,0) = -\frac{1}{c}\int\frac{\mathbf{K}^0}{r}\,dS - \frac{1}{c}\int\frac{\mathbf{K}^1}{r}\,dS - \frac{1}{c}\int ik\mathbf{K}^0\,dS + O(k^2).$$

The third integral on the right is of order k but it does not depend on x or y so that differentiation of this term with respect to x or y gives zero. In addition, $k^3\mathbf{F}$ is negligible compared with $\Delta\mathbf{F}$ in our approximation. If we then let

$$\mathbf{F}^0 = -\frac{1}{c}\int \frac{\mathbf{K}^0}{r}\,dS, \qquad \mathbf{F}^1 = -\frac{1}{c}\int \frac{\mathbf{K}^1}{r}\,dS, \qquad \ldots\ldots(9.9)$$

and expand all quantities in eqns. (9.6) to (9.8) in powers of k and compare corresponding terms, we ultimately obtain the differential equations

$$\left.\begin{aligned} &\frac{\partial F_y{}^0}{\partial x} - \frac{\partial F_x{}^0}{\partial y} = -\cos\phi_0 \sin\theta_0, \\ &\Delta F_x{}^0 = 0, \quad \Delta F_y{}^0 = 0, \end{aligned}\right\} \qquad \ldots\ldots(9.10)$$

and

$$\left.\begin{aligned} &\frac{\partial F_y{}^1}{\partial x} - \frac{\partial F_x{}^1}{\partial y} = -ikx\cos\phi_0 \sin^2\theta_0, \\ &\Delta F_x{}^1 = -ik\sin\phi_0\cos\theta_0, \quad \Delta F_y{}^1 = ik\cos\phi_0\cos^2\theta_0, \end{aligned}\right\} \qquad \ldots\ldots(9.11)$$

which, in combination with eqns. (9.10), constitute the differential-integral equations for the zeroth- and first-order terms $\mathbf{K}^0$ and $\mathbf{K}^1$, respectively, of the current density.

Note that if we had replaced the right-hand sides of eqns. (9.3) to (9.5) by their corresponding constant values at the centre we would have obtained eqns. (9.10) and (9.11) except for the first of eqns. (9.11) where the right-hand side would have become zero instead of $-ikx\cos\phi_0\sin^2\theta_0$. The theories of Bethe and Copson therefore can only lead to a correct $\mathbf{K}^0$ term; the term $\mathbf{K}^1$ is necessarily wrong in their approximation.

The solution of the differential-integral equations derived for $\mathbf{K}^0$ is not at all unique. In fact, it has been shown elsewhere (Bouwkamp 1953) that a seven-parameter family of solutions can be constructed, viz.

$$-\frac{1}{c}K_x{}^0 = \frac{p_0 + p_1x + p_2y - 2q_3xy + p_3(x^2 - y^2)}{\pi^2(a^2 - x^2 - y^2)^{1/2}},$$

$$-\frac{1}{c}K_y{}^0 = \frac{q_0 + q_1y + q_2x - 2p_3xy + q_3(y^2 - x^2)}{\pi^2(a^2 - x^2 - y^2)^{1/2}},$$

where the p's and q's are independent constants except for the relation $p_2 - q_2 = 2\cos\phi_0\sin\theta_0$. It is clear that additional conditions are necessary in order to find a unique solution. Of course, the charge density should be integrable over the aperture. In our case, this means that $xK_x{}^0 + yK_y{}^0$ should contain a factor $a^2 - x^2 - y^2$. Equivalently, the radial component $K_\rho{}^0$ of $\mathbf{K}^0$ must vanish at $\rho = a$. Using this condition we are actually left with a unique solution, namely,

$$\frac{1}{c}K_x{}^0 = -\frac{(\cos\phi_0 \sin\theta_0)y}{\pi^2(a^2 - x^2 - y^2)^{1/2}}, \qquad \frac{1}{c}K_y{}^0 = \frac{(\cos\phi_0 \sin\theta_0)x}{\pi^2(a^2 - x^2 - y^2)^{1/2}}.$$

In vector notation, this solution reads

$$\frac{1}{c}\mathbf{K}^0 = \frac{1}{2\pi^2}\frac{\boldsymbol{\rho} \times \mathbf{E}_0(0, 0, 0)}{(a^2 - \rho^2)^{1/2}}, \qquad \ldots\ldots(9.12)$$

which is exactly identical with Bethe's expression K_B. Note that $\operatorname{div}\mathbf{K}^0 = 0$, so that the corresponding charge density is zero.

The differential-integral equations for $\mathbf{K}^1$ can be solved in exactly the same way (Bouwkamp 1953). Let it suffice to quote here the final result in vector notation, applicable to arbitrary plane-wave excitation:

$$\frac{1}{c}\mathbf{K}^1 = \frac{1}{3\pi^2}\Big[(a^2-\rho^2)^{1/2}\{2ik\mathbf{H}_0 + \mathbf{n}\times\nabla(\mathbf{n}\,.\,\mathbf{E}_0)\} + \boldsymbol{\rho}\,.\,\{ik\mathbf{H}_0\times\mathbf{n} - \nabla(\mathbf{n}\,.\,\mathbf{E}_0)\}\frac{\mathbf{n}\times\boldsymbol{\rho}}{(a^2-\rho^2)^{1/2}}\Big]. \qquad \ldots\ldots(9.13)$$

Here $\mathbf{n}$ is the unit vector in the positive x direction and $\boldsymbol{\rho}$ the radius vector in the plane of the aperture while $\nabla(\mathbf{n}\,.\,\mathbf{E}_0)$ should be evaluated at the centre.

Bethe's expression (his K_H) is radically different from the correct one, viz.

$$\left(\frac{1}{c}\mathbf{K}^1\right)_{\text{Bethe}} = \frac{ik}{\pi^2}(a^2-\rho^2)^{1/2}\mathbf{H}_0.$$

The difference between our and Bethe's expressions for $\mathbf{K}^1$ is a solenoidal vector. Hence, Bethe's expression for the magnetic charge density is correct:

$$\eta = -\frac{1}{\pi^2}\frac{\boldsymbol{\rho}\,.\,\mathbf{H}_0}{(a^2-\rho^2)^{1/2}}. \qquad \ldots\ldots(9.14)$$

As we have seen before, the tangential magnetic and normal electric fields in the aperture are identical with the corresponding quantities of the undisturbed incident field. The remaining components in the aperture, in cylindrical coordinates, are

$$E_\rho = -\frac{2\cos\phi_0\sin\theta_0}{\pi}\frac{\rho}{(a^2-\rho^2)^{1/2}} - \frac{8ik}{3\pi}\frac{1}{(a^2-\rho^2)^{1/2}}[\sin\phi_0\cos\theta_0(a^2-\tfrac{1}{2}\rho^2)\sin\phi + \cos\phi_0\{(a^2-\tfrac{1}{2}\rho^2)+(\rho^2-\tfrac{1}{2}a^2)\sin^2\theta_0\}\cos\phi] + O(k^2a^2),$$

$$E_\phi = \frac{8ik}{3\pi}(a^2-\rho^2)^{1/2}[-\sin\phi_0\cos\theta_0\cos\phi + \cos\phi_0(1-\tfrac{1}{2}\sin^2\theta_0)\sin\phi] + O(k^2a^2),$$

$$H_\rho = \frac{4\rho}{\pi(a^2-\rho^2)^{1/2}}[\sin\phi_0\cos\theta_0\cos\phi - \cos\phi_0\sin\phi] + O(ka),$$

which expressions clearly illustrate the typical behaviour of the field vectors at the edge of the aperture. For the case of normal incidence ($\theta_0 = 0$, also $\phi_0 = 0$) this result reduces to the first few terms of an earlier one (Bouwkamp 1950f).

Whereas Bethe's expressions for the electric field in the aperture are radically different from the correct ones, his equations for the distant field (cf. Rayleigh 1897b) need not be modified. A simple explanation of this is due to Simon (1951) in connection with the fact that Bethe's $\mathbf{K}$ is correct except for the addition of a solenoidal vector. In accordance with Bethe the first-order diffracted field far behind the aperture can be represented as the superposition of an electric dipole field and a magnetic dipole field due to an electric dipole (moment $\mathbf{P}_e$) and a magnetic dipole (moment $\mathbf{P}_m$) both located at the centre of the aperture:

$$\mathbf{P}_e = +\frac{a^3}{3\pi}\mathbf{E}_0, \qquad \mathbf{P}_m = -\frac{2a^3}{3\pi}\mathbf{H}_0, \qquad \ldots\ldots(9.15)$$

directed along the normal and in the plane of the aperture, respectively. Note that the sign of $\mathbf{P}_e$ is incorrect in Bethe's paper.

The diffraction cross section of the small circular aperture for a plane wave polarized perpendicular to the plane of incidence is

$$\sigma_\perp = \frac{64}{27\pi} a^6 k^4 \cos^2\theta_0, \qquad \ldots\ldots(9.16)$$

and that for polarization in the plane of incidence is

$$\sigma_\parallel = \frac{64}{27\pi} a^6 k^4 [1 + \tfrac{1}{2}\sin^2\theta_0], \qquad \ldots\ldots(9.17)$$

which are special cases of the general expression

$$\sigma = \frac{64}{27\pi} a^6 k^4 [1 + (\tfrac{1}{2}\cos^2\phi_0 - \sin^2\phi_0)\sin^2\theta_0], \qquad \ldots\ldots(9.18)$$

valid for an arbitrary linearly polarized incident plane wave. For unpolarized incident waves (natural 'light') we get

$$\sigma' = \frac{64}{27\pi} a^6 k^4 [1 - \tfrac{1}{4}\sin^2\theta_0]. \qquad \ldots\ldots(9.19)$$

Equations (9.16) through (9.19) are due to Bethe (1944).

A few words must be said about the field near the aperture (the quasi-static field). We then have to calculate $\mathbf{F}$ for small values of ka and kr. This is most effectively done in spheroidal coordinates. Very tedious and complicated calculations are required to calculate the corresponding field. For final results the reader is referred to the author's report (Bouwkamp 1953).

By a systematic use of complicated integrals (Bouwkamp 1950c) Bethe's theory (for the case of normal incidence) has been extended to higher-order approximations (Bouwkamp 1950f). Six terms of the power-series solution have been derived. The fields on the obstacle and in the wave zone were given to this same degree of accuracy. The plane-wave transmission coefficient for the case of normal incidence was found to be

$$t = \frac{64}{27\pi^2}(ka)^4\left[1 + \frac{22}{25}(ka)^2 + \frac{7312}{18375}(ka)^4 + \ldots\right]. \qquad \ldots\ldots(9.20)$$

The same result was recently obtained by Flammer (1952) using spheroidal wave functions. The first two terms of eqn. (9.20) were first given by Meixner and Andrejewski (1950) and later on re-derived by Levine and Schwinger (1950) by means of the variational technique. Diffraction of a plane electromagnetic wave by a small conducting oblate spheroid of finite thickness for the case of normal incidence has been studied by Tai (1952b). When the eccentricity of the spheroid approaches zero Tai's solution reduces to the exact solution for a small sphere, and when the spheroid degenerates into a disc the expressions for the charge and current densities agree with the leading terms of the corresponding expressions obtained by Meixner and Andrejewski (1950) and Bouwkamp (1950f). For measurements on 'small' apertures, see Ehrlich (1951) or Silver, Ehrlich and Held (1952).

The theory of diffraction by small holes and discs has found important application in microwave technique (cf. Motz 1951, chap. 8). First of all we have to mention in this connection two reports by Bethe (1943a, b). Related work is due to Bernier (de Broglie 1948, p. 77), Casimir (1951), Chu and Hansen (1949), Dunning and Fellers (1951), Ginzton and Goodwin (1950), Klein (1951), Levin (1948b), Marcuvitz (1951a), Stevenson (1948), and Teichmann and

Wigner (1953). The effective permeability of an array of thin conducting discs was investigated by Estrin (1950) and the problems discussed by Ashour (1950, 1952) have also a bearing on our subject. Finally, an *experimental* method of determining the electric and magnetic polarizabilities of apertures and discs (Cohn 1951, 1952) significantly increases the utility of Bethe's small-aperture-coupling theory in the design of microwave devices.

§ 10. Mathieu Functions and Spheroidal Wave Functions in Diffraction Theory

As is well known, separation of the scalar wave equation in elliptic-cylindrical coordinates leads to the differential equation of Mathieu. Accordingly, diffraction problems in connection with an infinite elliptic cylinder (including the special cases of slits and ribbons) can be solved in terms of *Mathieu functions*. Similarly, separation in oblate and prolate spheroidal coordinates leads to *spheroidal wave functions*, which serve a useful purpose in the solution of diffraction problems relating to rotationally symmetric ellipsoids (so-called spheroids). In recent years these two types of function have attracted considerable attention in pure and applied mathematics, and a survey of the ever-increasing literature dealing with these special functions of mathematical physics seems appropriate.

An excellent treatise on the theory and application of Mathieu functions is due to McLachlan (1947), while extensive tables relating to Mathieu functions have recently been issued by the Computation Laboratory of the National Applied Mathematics Laboratories of the National Bureau of Standards, U.S.A. (1951). Earlier but less extensive numerical computations have been recorded by Stratton, Morse, Chu and Hutner (1941), who also gave tables for spheroidal wave functions. Additional tables of the latter functions were published by Bouwkamp (1947 b), for functions of order zero only. A new book on the theory of Mathieu functions and spheroidal wave functions by Meixner and Schäfke (1954) should prove invaluable to workers in the field. Original contributions to the theory and the numerical computation, asymptotic and series representation, etc., of these functions may be found in the papers by Abramowitz (1949), Bouwkamp (1948, 1950 b, d), Eberlein (1948), Flammer (1951), Leitner and Spence (1950 a), Meixner (1947, 1948 a, 1949 a, c, d, e, 1950, 1951 a, b), and Rydbeck (1945).

The available numerical tables relating to Mathieu functions and spheroidal wave functions permit the evaluation of the series solutions for the slit and circular-aperture problems up to kb or ka equal to 10 ($2b$ and $2a$ are the diameters of the slit and the circular aperture respectively), not so much for the whole field patterns as for the fields in the aperture and the wave zone as well as such overall quantities as the transmission cross section. In view of the further computational efforts at the National Bureau of Standards and the Massachusetts Institute of Technology, with the aid of high-speed computation machinery, it may soon be possible to calculate the diffraction phenomena referred to over a much wider range of parameters.

The first calculations on diffraction of plane waves by slits and ribbons, using Mathieu functions, were performed by Strutt (1931) and by Morse and Rubenstein (1938) for relatively small values of kb. Similar calculations with application to the theory of room acoustics were carried out by Pellam (1940).

In this paper a uniform strip of absorbing material in an otherwise totally reflecting plane wall served as a simple model to account for certain acoustic properties of a room. The diffraction effects of the strip were computed and the effect of the diffraction on the absorbing properties of the strip obtained. The radiation impedance of ribbon-type sound radiators was calculated by Nimura and Shibayama (1951). Miles (1949 b) and Dörr (1952) have pointed out that the two-dimensional (differential-) integral equations (2.25) and (2.26) can be solved in terms of Mathieu functions. Other theoretical work on Mathieu functions in diffraction theory is to be found in a book by Kupradze (1935), recently translated from the Russian. Two recent papers deserve special attention in connection with numerical results on diffraction by ribbons and slits. In the first, Moullin and Phillips (1952), using Morse and Rubenstein's expansions for the diffracted field, were particularly concerned with the numerical calculation of the electric currents induced in the perfectly conducting ribbon by a plane wave polarized parallel to the edge. As they emphasized, the distribution of induced currents near a bounding edge is a matter of great interest and is likely to serve as a guide in other problems not yet solved. However, these authors were apparently unaware of the existing literature on the edge singularities. Basing themselves on the early tables of Ince, they computed current distributions for $kb=2$ and $kb=4$ for several directions of incidence, including the cases of normal and of grazing incidence. Their discussion of very narrow ribbons seems somewhat complicated in view of the results quoted in §8 of the present Report. Also, many of their conclusions seem to require much more numerical evidence before any definite statements can be made. Discrepancies with Morse and Rubenstein's results are noted (but very unlikely due to errors in Mathieu-function tables as suggested). The second important paper is that of Skavlem (1951), who calculated the eigenvalues and eigenfunctions of the slit problem independently of any other worker (even the term Mathieu function does not occur). Skavlem's integral representation of the 'radial' Mathieu function is particularly noteworthy and analogous to one known in the theory of spheroidal wave functions (Bouwkamp 1941). Table 1 shows (rounded) values of the transmission coefficients (τ_1 and τ_2; see §8) computed by Skavlem for the case of normal incidence.

Table 1. Plane-Wave Transmission Coefficients (normal incidence) of an Infinite Slit (total width $2b$) according to Skavlem (1951)

kb	τ_1	τ_2	kb	τ_1	τ_2
0	0		2·0	1·1843	0·9948
0·2	0·0026		2·4	1·0865	
0·24		1·3965	3·0	0·9720	
0·4	0·0239		3·4	0·9282	
0·48		1·1216	4	0·9424	0·9990
0·6	0·0948		5	1·0499	
0·8	0·2606	1·0143	6	0·9956	
1·0	0·5454	0·9908	7	0·9717	
1·2	0·8769		8	1·0233	0·9998
1·4	1·1172	0·9813	9	1·0020	
1·6	1·2167		10	0·9822	
1·8	1·2213		∞	1	1

τ_1, electric vector parallel to the edge; τ_2, electric vector perpendicular to the edge; $k=2\pi/\text{wavelength}$.

Great activity has been displayed in recent years in the application of spheroidal wave functions to diffraction problems relating to the circular disc or the circular aperture, both for scalar diffraction (Bouwkamp 1941, Leitner 1949, Nimura 1951, Nimura and Watanabe 1953, Spence 1948 a, b, 1949, Spence and Leitner 1948, Storruste 1948, Storruste and Wergeland 1948 a, b, Wiener 1949 b, 1950, 1951) and for electromagnetic diffraction (Andrejewski 1951, 1952, Flammer 1952, Meixner 1946 a, 1948 b, Meixner and Andrejewski 1950). In most applications the simple case of a normally incident plane wave is treated Then, in the scalar theory, only spheroidal wave functions of order zero are involved, while in the electromagnetic case functions of orders zero and two appear in the analysis. In this connection it may be remarked that certain quasi-vectorial problems (Bardeen 1930, Leitner 1950, Leitner and Spence 1950 b, Meixner 1953) can be solved in terms of spheroidal wave functions of order one (see also Storer 1951, 1952). For obstacles of finite thickness (Spence and Granger 1951), for a dipole incident field (Meixner 1953), and for obliquely incident waves the analysis becomes much more complicated, so that numerical evaluation of the formal solutions has not yet been accomplished to any significant extent.

We shall now discuss in some detail the diffraction of a scalar plane wave by a circular aperture for the case of normal incidence. Let ξ, η, ϕ be oblate spheroidal coordinates, related to the cylindrical coordinates ρ, ϕ, z by the transformation

$$\rho/a = [(1-\xi^2)(1+\eta^2)]^{1/2}, \qquad z/a = \xi\eta, \qquad (\eta \geqslant 0, \quad -1 \leqslant \xi \leqslant 1)$$

where a is the radius of the aperture. In virtue of the axial symmetry of the problem, we can expand the wave functions ϕ_1 and ϕ_2 (for their definition see § 2; the incident wave is represented by $u_0 = \exp(ikz)$) in terms of spheroidal wave functions of order zero, viz.

$$\phi_{1,2} = \Sigma_\nu \alpha_\nu X_\nu(\xi) Y_\nu(\eta)$$

where the coefficients α_ν must be determined from the boundary conditions. The 'angular' spheroidal wave functions $X_\nu(\xi)$ form a complete and orthogonal set of eigenfunctions on the interval $-1 \leqslant \xi \leqslant 1$ of either the *differential equation*

$$(1-\xi^2)X'' - 2\xi X' + (\lambda + k^2a^2\xi^2)X = 0 \qquad \ldots\ldots(10.1)$$

or the *integral equation*

$$X(\xi) = \mu \int_{-1}^{1} \exp(ka\xi t) X(t)\, dt. \qquad \ldots\ldots(10.2)$$

The eigenvalues λ_ν of the differential equation (10.1) and the eigenvalues μ_ν of the integral equation (10.2) depend of course on the parameter ka and the degree ν ($\nu = 0, 1, 2, \ldots$). The angular function $X_\nu(\xi)$ is an even or odd function of ξ according as ν is even or odd; in addition it is an entire function of ξ, with no singularities in the finite ξ plane. The normalization is chosen so as to make

$$\int_{-1}^{1} \{X_\nu(\xi)\}^2\, d\xi = \frac{2}{2\nu+1},$$

independently of the value of ka. In the limiting case $ka \to 0$, $X_\nu(\xi)$ coincides with $P_\nu(\xi)$, the Legendre polynomial of degree ν.

The 'radial' function $Y_\nu(\eta)$ is conveniently defined by an integral representation in terms of the corresponding 'angular' function

$$Y_\nu(\eta) = \tfrac{1}{2}\, e^{ika\eta} \int_{-1}^{1} \frac{e^{-ka\xi} X_\nu(\xi)}{i\eta - \xi}\, d\xi. \qquad \ldots\ldots(10.3)$$

This function may be considered as a generalization of the associated Legendre function $Q_\nu(i\eta)$, obtained from eqn. (10.3) by letting ka tend to zero.

The two boundary-value problems discussed in §2, for the special case of a circular aperture, are solved by the expansions

$$\phi_1 = \sum_{\nu=0}^{\infty} \frac{2\nu+\frac{3}{2}}{\mu_{2\nu+1}} \frac{iX'_{2\nu+1}(0)}{Y'_{2\nu+1}(0)} X_{2\nu+1}(\xi) Y_{2\nu+1}(\eta), \qquad \text{.....(10.4)}$$

and

$$\phi_2 = \sum_{\nu=0}^{\infty} \frac{2\nu+\frac{1}{2}}{\mu_{2\nu}} \frac{X_{2\nu}(0)}{Y_{2\nu}(0)} X_{2\nu}(\xi) Y_{2\nu}(\eta). \qquad \text{.....(10.5)}$$

The coefficients occurring in these expansions can be written in terms of the eigenvalues μ_ν, since

$$\frac{Y_{2\nu}(0)}{X_{2\nu}(0)} = \frac{ka}{(2\mu_\nu)^2} - \frac{\pi}{2}i, \qquad \frac{iY'_{2\nu+1}(0)}{X'_{2\nu+1}(0)} = \frac{ka}{(2\mu_{2\nu+1})^2} + \frac{\pi}{2}i,$$

so that

$$\phi_1 = -\sum_{\nu=0}^{\infty} \frac{2(4\nu+3)\mu_{2\nu+1}}{ka+2\pi i(\mu_{2\nu+1})^2} X_{2\nu+1}(\xi) Y_{2\nu+1}(\eta),$$

$$\phi_2 = \sum_{\nu=0}^{\infty} \frac{2(4\nu+1)\mu_{2\nu}}{ka-2\pi i(\mu_{2\nu})^2} X_{2\nu}(\xi) Y_{2\nu}(\eta).$$

Using the asymptotic relations

$$Y_\nu(\eta) = (-1)^\nu \frac{e^{ika\eta}}{2\mu_\nu i\eta} X_\nu(1)\{1+O(\eta^{-1})\} \qquad \text{.....(10.6)}$$

and $a\eta \sim r$, $\xi \sim \cos\theta$, where r and θ are spherical polar coordinates around the axis of the circular aperture, we find for the distant field ($r \to \infty$)

$$\phi_1 \sim \frac{e^{ikr}}{ikr} \sum_{\nu=0}^{\infty} \frac{(4\nu+3)X_{2\nu+1}(1)}{1+(2\pi i/ka)(\mu_{2\nu+1})^2} X_{2\nu+1}(\cos\theta), \qquad \text{.....(10.7)}$$

$$\phi_2 \sim \frac{e^{ikr}}{ikr} \sum_{\nu=0}^{\infty} \frac{(4\nu+1)X_{2\nu}(1)}{1-(2\pi i/ka)(\mu_{2\nu})^2} X_{2\nu}(\cos\theta), \qquad \text{.....(10.8)}$$

and for the transmission coefficients

$$t_1 = \frac{4}{(ka)^2} \sum_{\nu=0}^{\infty} \frac{(2\nu+\frac{3}{2})[X_{2\nu+1}(1)]^2}{1+(2\pi/ka)^2(\mu_{2\nu+1})^4}, \qquad \text{.....(10.9)}$$

$$t_2 = \frac{4}{(ka)^2} \sum_{\nu=0}^{\infty} \frac{(2\nu+\frac{1}{2})[X_{2\nu}(1)]^2}{1+(2\pi/ka)^2(\mu_{2\nu})^4}. \qquad \text{.....(10.10)}$$

These expansions are somewhat simplified if we use a notation suggested by Storruste (1948) by introducing the auxiliary angle δ_ν defined by

$$\tan\delta_\nu = ka/2\pi\mu_\nu^2. \qquad \text{.....(10.11)}$$

Then eqns. (10.7) and (10.8) become

$$\phi_1 \sim \frac{e^{ikr}}{ikr} \sum_{\nu=0}^{\infty} (2\nu+\tfrac{3}{2})\{1-\exp(2i\delta_{2\nu+1})\}X_{2\nu+1}(1)X_{2\nu+1}(\cos\theta), \qquad \text{.....(10.12)}$$

$$\phi_2 \sim \frac{e^{ikr}}{ikr} \sum_{\nu=0}^{\infty} (2\nu+\tfrac{1}{2})\{1-\exp(-2i\delta_{2\nu})\}X_{2\nu}(1)X_{2\nu}(\cos\theta), \qquad \text{.....(10.13)}$$

and the simplified equations for the transmission coefficients are

$$t_1 = \left(\frac{2}{ka}\right)^2 \sum_{\text{odd } n \geqslant 1} (n+\tfrac{1}{2}) \sin^2\delta_n \{X_n(1)\}^2, \qquad \text{.....(10.14)}$$

$$t_2 = \left(\frac{2}{ka}\right)^2 \sum_{\text{even } n \geqslant 0} (n+\tfrac{1}{2}) \sin^2\delta_n \{X_n(1)\}^2. \qquad \text{.....(10.15)}$$

To facilitate comparison with Storruste's results, we mention that his parameters $\eta_{2\nu+1}$, $-\delta_{2\nu}$ are identical with our parameters $\delta_{2\nu+1}$, $\delta_{2\nu}$ respectively. Storruste's equation for ϕ_1 has the wrong algebraic sign as is easy to verify in connection with Levine and Schwinger's cross section theorem, eqn. (2.22). An invariant characterization of the parameter δ_ν is contained in the statement that the angular spheroidal wave functions of order zero are the eigenfunctions of the integral equation

$$\lambda X(\xi) = \frac{1}{\pi \tan \delta} \int_{-1}^{1} \frac{\sinh [ka(\xi + t)]}{\xi + t} X(t)\, dt$$

with $\delta = \delta_\nu$ as the eigenvalues.

In table 2 we have presented some numerical values of the transmission coefficients, based on Storruste's and our own numerical calculations.

Table 2. Numerical Values of the Plane-Wave Transmission Coefficients of a Circular Aperture in a Perfectly Soft Screen (t_1) and a Perfectly Rigid Screen (t_2) for the case of Normal Incidence

ka	t_1	ka	t_1	ka	t_1	ka	t_1	ka	t_1
0·1	0·0000030(	1·1	0·064	2·1	1·042	3·1	1·109	4·1	0·880
0·2	0·0000486	1·2	0·097	2·2	1·135	3·2	1·078	4·2	0·872
0·3	0·000250	1·3	0·142	2·3	1·200	3·3	1·048	4·3	0·867
0·4	0·000809	1·4	0·203	2·4	1·235	3·4	1·020	4·4	0·866
0·5	0·00203	1·5	0·283	2·5	1·248	3·5	0·994	4·5	0·869
0·6	0·00436	1·6	0·384	2·6	1·244	3·6	0·969	4·6	0·878
0·7	0·00843	1·7	0·505	2·7	1·227	3·7	0·946	4·7	0·895
0·8	0·0151	1·8	0·642	2·8	1·202	3·8	0·926	4·8	0·915
0·9	0·0255	1·9	0·783	2·9	1·173	3·9	0·908	4·9	0·937
1·0	0·0411	2·0	0·922	3·0	1·142	4·0	0·893	5·0	0·958

ka	t_2	ka	t_2	ka	t_2	ka	t_1	t_2
0	0·8106	1·2	0·854	3·4	0·986	6·0	1·065	0·993
0·1	0·8109	1·4	0·868	3·6	0·988	7·0	0·964	
0·2	0·8118	1·6	0·883	3·8	0·988	8·0	0·967	0·996
0·3	0·8134	1·8	0·900	4·0	0·987	9·0	1·039	
0·4	0·8156	2·0	0·917	4·2	0·987	10·0	0·987	0·998
0·5	0·8185	2·2	0·932	4·4	0·986	∞	1	1
0·6	0·8219	2·4	0·945	4·6	0·986			
0·7	0·8260	2·6	0·958	4·8	0·987			
0·8	0·8306	2·8	0·968	5·0	0·988			
0·9	0·8358	3·0	0·976					
1·0	0·8415	3·2	0·982					

We now turn to the analogous diffraction problem in electromagnetic diffraction theory, that is, the diffraction of a plane polarized electromagnetic wave by a perfectly conducting circular disc of negligible thickness for the case of normal incidence. As has been first realized by Bouwkamp (1941), Möglich's (1927) solution of this problem is erroneous because of the inadmissible high singularities at the edge (Bouwkamp 1946 b). Also in error is Meixner's (1946 a) first attempt to solve this fascinating problem. Yet the first correct solution was given by Meixner (1948 b) who then realized that the assumption of homogeneous boundary conditions for the Hertzian (Möglich 1927) or the Debye potentials (Meixner 1946 a) necessarily led to an incorrect solution, violating the

edge condition. Meixner's solution in terms of Debye potentials (Meixner 1948 b) is unsatisfactory in the sense that the series expansions involved are weakly convergent and that the Debye potentials are highly singular at the origin of coordinates. These singularities, however, do not lead to singularities in the field vectors themselves. A much simpler solution was found by Meixner and Andrejewski (1950), using the ordinary vector potentials (or equivalent Hertzian vectors, in the time-periodic case). It would be of some interest to prove in a direct manner that the two solutions (of Meixner, and of Meixner and Andrejewski) are identical. However, such a direct proof has not yet been given.

If the incident wave is polarized parallel to the x axis, it follows from Meixner and Andrejewski's theory that the two components of the current density in the circular disc are of the form

$$I_x = \frac{A(\rho) + B(\rho)\cos 2\phi}{(a^2-\rho^2)^{1/2}}, \qquad I_y = \frac{B(\rho)\sin 2\phi}{(a^2-\rho^2)^{1/2}}$$

where $A(\rho)$ and $B(\rho)$ are functions uniformly bounded for $0 \leqslant \rho \leqslant a$ and satisfying $A(a)+B(a)=0$, $B(0)=0$. Accordingly the diffracted field can be derived from a vector potential $\mathbf{A}$ with components

$$A_x = C(z,\rho) + D(z,\rho)\cos 2\phi, \qquad A_y = D(z,\rho)\sin 2\phi, \qquad A_z = 0,$$

where $C(z,\rho)$ and $D(z,\rho)$ may be expanded in oblate spheroidal wave functions of order zero and two respectively.

In his beautiful dissertation, Andrejewski (1951, 1952) has calculated the reflection coefficient of the disc in the range $0 \leqslant ka \leqslant 10$. The diffraction patterns of the field and Poynting vectors were computed for $ka = 10$. The fields and intensities on the disc (or in the complementary aperture) and along the symmetry axis were given for $ka = 10$, while the same quantities at the centre of the disc (or aperture) were calculated for $0 \leqslant ka \leqslant 10$. The theoretical values were compared with the experimental results of Andrews (1950 b) and of Severin (1946). To illustrate the significance of the edge condition, Andrejewski evaluated and compared Möglich's solution with the rigorous solution. In addition, all possible approximations discussed in the preceding sections of this Report were carefully checked against the rigorous solution, including Braunbek's theory (of § 4) which was extended to electromagnetic diffraction by Frahn (1951). Finally, a great number of tables pertaining to oblate spheroidal wave functions are included in Andrejewski's work.

§ 11. Diffraction by Spherical and Cylindrical Obstacles

An important, if not the most important, problem of diffraction is that in which a monochromatic plane wave impinges on a homogeneous isotropic sphere embedded in an infinite homogeneous medium. A formal solution of this, both for sound diffraction and for electromagnetic diffraction, was given about fifty years ago (Rayleigh, Mie, Debye) in the form of an infinite series involving Bessel functions of orders equal to half an odd integer. This series solution includes all possible cases as regards the physical constitution of the sphere in relation to its surrounding (absorbent, transparent, perfectly reflecting spheres). However, in many practical applications the numerical evaluation of the solution is laborious, if not impossible, owing to the slow convergence of the series which necessitates the inclusion of a large number of terms. In principle, there are

two possibilities to meet this difficulty. Either the series solution may be transformed by analytic means into others that are better convergent or possibly asymptotic in a certain range of parameter values, or a wider range of variables may be covered by an extensive tabulation of the various auxiliary functions that occur in the original series solution. In both respects, important progress has been achieved lately. This holds equally for diffraction by circular cylinders where Bessel functions of integral order occur.

In what follows, we use the notation α for the characteristic parameter $2\pi a/\lambda$, where λ is the wavelength in the external medium and a the radius of the sphere or cylinder. Further, m denotes the refractive index of the obstacle, that is, the ratio of wave velocities of medium to obstacle. Similarly μ is the density ratio.

Extending the work of King, Fox (1940) gave tables allowing the determination of the mean radiation pressure on a sphere in a plane progressive wave in the range $\alpha = 1(1)11$, 15, 20 and $\mu = 0$, 0·1, 0·2, 0·5, 1·0. A simple interpolation formula was derived which may also be used for extrapolation to very large α with accuracy better than one per cent. The maximum number of terms in the case of $\alpha = 20$ was 26. Diffraction of a plane sound wave by a perfectly rigid sphere was considered by Schwarz (1943). Accurate numerical tables for the pressure distribution in amplitude and phase at the surface of the sphere were given for $\alpha = 1(1)10$ and $\theta = 0(5)180°$ where θ denotes the polar angle. The maximum number of terms was 22. Schwarz's paper contains an excellent review of earlier results of Rayleigh, Bellantine, Stenzel, and others, while his own results are displayed in a number of curves. Schwarz's calculations were compared with experiments by Wiener (1947). On the whole, reasonably good agreement was obtained. In particular, the existence of the bright spot in the middle of the 'dark' face of the sphere was verified experimentally. Wiener also gave tables for the pressure distribution along the surface of an infinite rigid circular cylinder for the case of normal incidence and $\alpha = \frac{1}{2}$, $1(1)10$, $\theta = 0(10)180°$ where θ is the azimuthal angle. In addition, he compared experimental results for a finite cylinder (of height equal to diameter) with those for the sphere of equal size. Cook and Chrzanowski (1946) calculated the absorption and scattering of plane waves of sound for a cylinder of known normal acoustic impedance. The appropriate reverberation-room statistics were developed and applied. Similar calculations were carried out for spherical absorbers. In connection with the scattering of sound by marine life and the problem of sound ranging in the ocean, Anderson (1950) considered the sound scattering from a fluid sphere for values of μ and m between $\frac{1}{2}$ and 2. The reflectivity of the sphere and total-scattering curves were calculated up to $\alpha = 16$. Spheres with acoustic properties near those of the external medium were theoretically studied by Hart (1951). Under this restriction simple analytic expressions in closed form can then be obtained for the reflectivity and the scattering cross section. A typical example for the scattering cross section of a large sphere, valid when μ and m are near unity, is

$$\frac{\sigma}{2\pi a^2} \sim \frac{2m^2(1+\mu^2)}{(\mu m+1)^2}\left[1+\frac{m^2-1}{2m}\frac{\mu^2-1}{\mu^2+1}\log\frac{m+1}{m-1}\right]. \quad \ldots\ldots(11.1)$$

The author would remark that eqn. (11.1) is essentially wrong if μ and m are exactly equal to unity as in this case there is no scattered field at all. Presumably a further condition is that $(m-1)\alpha$ should not be too near zero. Two very useful

approximations to the scattering cross section were derived for the cases $\mu = 1$, $m \simeq 1$ and $m = 1$, $\mu \simeq 1$ as a function of α. Calculated values in the range $0 \leqslant \alpha \leqslant 16$ showed very good agreement with Anderson's exact values for $m = 5/4$, $5/6$, $\mu = 1$. Extensive tables of auxiliary functions in absorption and scattering problems for spheres and circular cylinders are now available (Lowan, Morse, Feshbach and Lax 1946, Lax and Feshbach 1948). They are based on the phase-shift analysis while the boundary condition at the obstacle is to be expressed in terms of an impedance. The range is $\alpha = 0{\cdot}0(0{\cdot}1)10{\cdot}0$. So far the spherical or cylindrical obstacle was assumed to be isotropic. Extension to scatterers of solid material involving shear waves in addition to compressional waves was considered by Faran (1951). The results can be expressed in terms of the scattering functions tabulated by Lowan, Morse, Feshbach and Lax. Calculated scattering patterns were shown to be in good agreement with experiments on water-embedded metal cylinders. Resonance phenomena were observed near frequencies of normal modes of the freely vibrating scatterer. Finally, there is a little-known paper of Wergeland (1945) on the diffraction of a plane wave of sound by a rigid sphere where the phase-shift analysis is applied for the calculation of the total scattering cross section in the range $0 \leqslant \alpha \leqslant 10$. There is a detailed discussion of the apparent paradox (cf. Jobst 1925) that the cross section tends to twice the geometric one as $\alpha \to \infty$. The scattered intensity can be separated into two parts. One part corresponds to a spherical wave of uniform intensity in all directions, the other is confined to a narrow region around the primary direction and is connected with the geometrical-optics shadow. A similar paradox is known in certain quantum-mechanical problems. Its electromagnetic equivalent was discussed by Van de Hulst (1946), Sinclair (1947), L. Brillouin (1949), and Sifrin (1950a).

As already remarked in § 1, numerical results for the diffraction of an electromagnetic wave by a spherical particle are useful in a great number of problems occurring in physics, chemistry, meteorology, and radio engineering. In what follows, we have omitted all reference to work that was reviewed earlier by Oster (1948), Van de Hulst (1946, 1949b), Bremmer (1949) or Burrows and Attwood (1949). Further, papers that have appeared later will be discussed only superficially. Various contributions to radio-wave propagation around the spherical earth will be found in the papers by Domb (1953), Friedman (1951a, b), Friedman and Russek (1952), Ljunggren (1948, 1949), Marcuvitz (1951b), Miller (1951), Nomura (1951), Pryce (1953), Rydbeck (1944), Schachenmeier (1951), Tai (1952c), and in the book of Sommerfeld (1948b).

Extensive tabular material is now available for the diffraction of a plane electromagnetic wave by a homogeneous spherical particle, either transparent or absorbent. In the first place, there are the *Tables of Scattering Functions for Spherical Particles* (National Bureau of Standards, Applied Mathematics Series, 4, Washington, 1949), giving the intensities of the two incoherent plane polarized components scattered by a particle illuminated with natural light in the range $0{\cdot}5 \leqslant \alpha \leqslant 6{\cdot}0$ for $m = 1{\cdot}33$, $1{\cdot}44$, $1{\cdot}55$, $2{\cdot}00$ and $\theta = 0(10)180°$. There is also a table for the total scattering coefficient; for $m = 1{\cdot}50$ up to $\alpha = 12$. Further tables deal with absorbent particles (complex values of the refractive index m); the extinction coefficient varies between 0 and 0·1 and the real refractive index from 1·44 to 1·55. Numerical data for certain selected values of the complex index are included, as well as many useful polynomial approximations for

interpolation purposes. Secondly, we have to mention the recent tables prepared by Gumprecht and Sliepcevich (1951c). These tables cover a fifty-times greater range of diameters, up to 127 wavelengths. Six (instead of four) real indices of refraction were chosen: $m = 1{\cdot}20, 1{\cdot}33, 1{\cdot}40, 1{\cdot}44, 1{\cdot}50, 1{\cdot}60$. On the other hand data relating to absorbent spheres are not given. Also, the intensity functions are tabulated for an angle of observation of 90° only. Auxiliary tables by the same authors (1951a, b) may be used for calculating the intensity functions in other directions. The range of α is given by $\alpha = 1(1)6(2)10(5)100(10)200(50)400$. The maximum number of terms of the Mie series taken into account was 421. The tables were computed on the ENIAC: "It is estimated that it would have taken one man, working 40 hours per week with a standard desk computer, approximately 25 years to make similar computations. Less than two weeks' time was required once the details of the programming had been worked out".

Numerous theoretical and experimental papers have appeared in the last few years. Values of the scattering cross section of water drops in air were computed by Houghton and Chalker (1949) on the assumption that water drops are dielectric spheres of refractive index $m = 4/3$. Fifty different values of α, ranging from 7 to 24, were considered. There is a discussion of the maxima and minima in the curve of the scattering coefficient versus α in connection with earlier work by Ryde (1946) and Van de Hulst (1946). Theissing (1950) studied light scattering by an infinite layer of suspended dielectric particles, using the old numerical results of Blumer. The first laboratory measurements of the back-scattering cross section on individual water spheres in the range $0{\cdot}74 \leqslant \alpha \leqslant 5{\cdot}90$ were made by Aden (1950, 1951a, b) for thirty different radii. Good agreement was obtained with theoretical values; measurements were checked with eighteen metal spheres. Similar measurements on diffraction by spheres and edges at $\lambda = 1{\cdot}25$ cm were performed by Huang and Kodis (1951) and Kodis (1952); for measurements on cylinders, see Kodis (1950). A theory on the scattering of plane waves by soft spherical particles was developed by Hart and Montroll (1951). By making suitable approximations to the Bessel-function terms in the exact Mie solution, these authors were able to derive analytical closed-form expressions for the total scattering cross section of the sphere and the intensity of the scattered wave. Comparison with the exact values revealed that the relative error in the approximate cross section is less than five per cent for scatterers of any size provided that $m < 1{\cdot}1$; when $1{\cdot}1 < m < 1{\cdot}5$, the relative error is less than five per cent if $\alpha(m-1) < 6$ and less than 25 per cent for any size of sphere. Similar work on cylinders, spheroids, and discs is also available (Montroll and Hart 1951, Montroll and Greenberg 1952). Some of the results were independently obtained by Van de Hulst (1946, 1950).

Other papers that deserve to be mentioned in this connection are those of Holl (1946, 1948) and Bucerius (1946), the latter dealing with the theory of the rainbow and the glory. Scattering of light by large drops of water and polarization of light in rainbows was discussed by Šlfrin (1950b); cf. also a recent paper by Lander and Nielsen (1953). Polarization of light diffused by small spherical particles was treated by Perrin and Abragam (1951). Light scattering by non-spherical particles, as spheroids and cylinders, whose dielectric constant is slightly different from that of the solvent, was studied by Saitô and Ikeda (1951). Scattering and attenuation by non-spherical particles in the

atmosphere was studied by Atlas, Kerker, and Hitschfeld (1953), while Kerker, Langleben and Gunn (1951) considered the scattering of microwaves by a melting spherical ice particle. Radiation from a circular loop antenna in the presence of a sphere or a spheroid was theoretically treated by Tai (1952d). The same author (Tai 1948) has also studied the reflection of plane electromagnetic waves from a perfectly conducting grounded half-cylinder. Finally, Aden and Kerker (1951) derived formulae for the scattering of electromagnetic waves from two concentric spheres, and Güttler (1952) studied the diffraction by a dielectric sphere with absorbing nucleus for applications to interstellar absorption of light. Other papers to be mentioned are those of Aden (1952) and Labrum (1952) dealing with back-scattering of electromagnetic waves from spheres and spherical shells and scattering of radio waves by meteorological particles respectively. Rževkin (1951) studied the connection of the problem of sound diffraction by a sphere with the reciprocity theorem.

Most papers so far discussed are concerned with the diffraction of plane waves. In short-wave radio propagation (cf. Bremmer 1949), however, the source of radiation is often taken in the form of a dipole, either electric or magnetic. For example, the currents induced on the surface of a conducting sphere excited by a radial dipole antenna carrying a sinusoidal current distribution were calculated by Papas and King (1948). Patterns of Hertzian dipoles and other antennae in the presence of circular or elliptic cylinders were displayed in great number by various authors including Carter (1943), Oberhettinger (1943), Lucke (1951a), and Sinclair (1951). Much numerical material in connection with the physical interpretation of the expression for an outgoing wave in cylindrical coordinates is to be found in a paper by Junger (1953). Franz and Deppermann (1952) considered the diffraction by a cylinder by introducing a type of surface wave. Scattering of cylindrical waves by a cylinder was studied by Faran (1953). Among the many papers that dealt with the radiation of slotted cylinders, and therefore closely related to diffraction studies, we mention those of Haycock and Wiley (1952), Laird and Cohen (1952), Papas (1949), Papas and King (1949), Rhodes (1950), Silver and Saunders (1950a, b), and Sinclair (1948). Reflection and transmission of sound and electromagnetic waves by thin curved shells has been studied in detail by Primakoff and Keller (1947), Greenstadt (1948), Keller and Keller (1949), Keller (1950), and Junger (1952). Reflection of a pulse by a spherical surface was treated by Anderson, Northwood and Barnes (1952), while Barnes and Anderson (1952) studied the sound field from a pulsating sphere and the development of a tail in pulse propagation. For a study on the determination of reflected and transmitted fields by geometrical optics, see Keller and Keller (1950) and Keller and Preiser (1950).

Several authors have dealt with the problem of *multiple* scattering. Diffraction of plane waves by two parallel cylinders, or by an arbitrary configuration of parallel cylinders, was studied in a series of papers by Twersky (1950, 1951a, b, 1952a, b, c). The variational technique was applied by Shmoys (1951) for the diffraction of electromagnetic waves by a plane wire grating (cf. also Sollfrey and Shmoys 1952). Reflection and transmission by wire gratings was also treated by Honerjäger (1949) and by Lewis and Casey (1952) who considered resistive wires. Related papers are those of Sakurai (1950a) and R. Müller (1953a). Details of any of these papers are beyond the scope of this Report.

§ 12. Survey of Other Problems in Diffraction Theory

In this section an attempt will be made to give the reader an impression of recent work that for one reason or another has not been incorporated into the preceding sections. Generally a few words must suffice to indicate the character of the problems treated. The relevant references are given as completely as possible so as to help the interested reader to find his way through the bulk of papers scattered throughout the many journals. Lack of space (and time) prevents us from giving details.

Artmann (1950) has studied the diffraction of polarized light by screens of finite thickness in the shadow zone, while Born (1943) and Braunbek (1949) were concerned with the distribution of energy near diffracting discs, mirrors, and edges. Diffraction of radar waves by a semi-infinite screen was studied by Horton (1949a), Horton and Watson (1950), Watson and Horton (1950), and Harden (1952). Experimental results of the latter author for sheets of different thicknesses and conductivities show the importance of these factors in the field near the diffracting edge. Booker and Clemmov (1950b) discussed the relation between the Sommerfeld theory of radio propagation over a flat earth and the theory of diffraction at a straight edge. Calculations on diffraction from an irregular screen were applied to ionospheric problems by Booker, Ratcliffe and Shinn (1950). McDonald and Harris (1952) considered the diffraction of spherical scalar waves, Senior (1953) the diffraction of dipole waves, and Clemmov (1950b) the diffraction of cylindrical waves, by a perfectly conducting half-plane.

For diffraction of sound pulses by semi-infinite planes, by wedges, or by corners the reader is referred to Berry (1952b), Friedlander (1946), Keller and Blank (1951), and Miles (1952b). The latter author (Miles 1952c) considered also the case of electromagnetic pulses. Further contributions to diffraction by wedges and corners are due to Felsen (1952), Garnir (1952a, b), Herglotz (1952), Jones and Pidduck (1950), Karp and Sollfrey (1950), Kay (1952), Keller (1952), and Sollfrey (1952). Other papers on diffraction by half-planes and related topics are those of Berry (1952a), Braunbek and Laukien (1952), Clemmov (1950a), Friedlander (1951), Harrington (1953), Jones (1950, 1952d), Maréchal and Naudin (1950), Savornin (1939), Savornin and Laurentjoye (1946), and Senior (1952).

Sound diffraction by rigid circular cones was studied by Wiener (1948), electromagnetic scattering from wedges and cones by Sletten (1952), while the electromagnetic field radiated by the conical horn was theoretically investigated by Schorr and Beck (1950). A new analytical approach to wave problems in connection with conical obstacles (Kontorowich and Lebedev 1939, Lebedev 1949) deserves special attention. Chester (1952) studied the reflection of a transient pulse by a parabolic cylinder and a paraboloid of revolution (see also Anderson 1952), while Horton and Karal (1950) treated the diffraction of a plane sound wave by a paraboloid of revolution (cf. Horton 1953). Buchholz (1942) has determined the field produced by a point source located at the focus of the same paraboloid, as well as the axially symmetric radiation between confocal paraboloids of revolution under various conditions of excitation (Buchholz 1948a). Similar electromagnetic problems are treated in the papers by Horton and Karal

(1951), Mirimanov (1948 a, b), and Pinney (1947). The relevant analysis involves Laguerre functions, which were studied in detail by the same authors (Buchholz 1943 a, 1948 b, 1953, Pinney 1946). An earlier contribution to the theory of the cylindrical-parabolic mirror is due to Magnus (1941 b).

Interesting approximations in diffraction theory with applications to conducting obstacles of general shape have been discussed in a series of papers by the Russian scientist Fock (1946 a, b, 1948, 1950, 1951).

Several authors have recently been concerned with problems of uniqueness, existence, and similar more fundamental questions in relation to diffraction theory. Let it be sufficient to cite here the papers by Freudenthal (1939), Hönl and Maue (1952), Kupradze (1939), Ledinegg and Urban (1952), Maue (1952), Magnus (1942, 1949), Magnus and Rellich (1943), C. Müller (1949 a, b, 1950, 1951), Rellich (1943), Saunders (1950, 1951, 1952), and Weyl (1952 a, b).

The following list of papers contains recent contributions to the theory of diffraction by obstacles and slots in waveguides, and the radiation of open-ended waveguides: Buchholz (1943 b, 1948 c), Dockès (1952), Feuer and Akeley (1948), Levin (1948 a), Lewin (1949, 1952), Magnus and Oberhettinger (1950), Marcuvitz and Schwinger (1951), Miles (1946 b, 1947 a), R. Müller (1952, 1953 b), Ruch (1950) and Zernov (1951 b).

Several elementary considerations on Fresnel diffraction, Fraunhofer diffraction, radiation from edges, etc., may be found in the papers of Bouwkamp (1940), Faust (1951), Hartnagel and Kappler (1949), Kathavate (1945 a), Lamar (1949), Pidduck (1946, 1947, 1951), and Ramachandran (1945). Blanc-Lapierre and others (cf. Stumpers 1953, p. 39) have studied diffraction phenomena from the point of view of information theory. A photographic method for displaying sound waves and microwave patterns was developed by Hubbard, Zartman, and Larkin (1947) and by Kock and Harvey (1951). Applications of vector analysis to the wave equation were made by Miles (1948 b) and Spence and Wells (1951).

Several papers on the diffraction by slits and circular apertures or related problems may be mentioned. First there is the paper by Van Vleck, Bloch and Hamermesh (1947) in which radar reflection from wires or thin metallic strips (dropped in large quantities over continental Europe during the second world war) was studied in detail, using modern techniques of integral-equation approach (Hallén) to antenna-radiation problems. In a closely related paper of Levin (1944), resonance effects of a slit of finite length were discussed. Similar work is due to Zernov (1951 a). Awatani (1951), Kawai (1952), and Sasao (1932), of which the second reference only is available to the author, dealt with acoustic diffraction by a circular aperture or a circular plate (Rayleigh disc). The problem of the circular diffraction antenna (Pistolkors 1948) was solved in terms of spheroidal wave functions by Kloepfer (1952); for a different approach, see Ludwig (1951) and Meixner and Kloepfer (1951). Kirchhoff-type approximations to the transmission coefficients of annular, elliptical, and rectangular apertures were evaluated for normal incidence by Huang (1953 a). A review and comparison of the various methods of attacking the problem of diffraction by a circular aperture is promised by the same author (Huang 1953 b). Finally, Anders (1953) considered the diffraction by a circular aperture from the point of view of dual integral equations; his analysis is similar to that of Sommerfeld (1942).

References

Abramowitz, M., 1949, *J. Math. Phys.*, **28**, 195.

Aden, A. L., 1950, *Electromagnetic Scattering from Metal and Water Spheres*, Cruft Laboratory, Technical Report No. 106; 1951 a, *J. Appl. Phys.*, **22**, 601; 1951 b, *Amer. J. Phys.*, **19**, 163; 1952, *Back-Scattering of Electromagnetic Waves from Spheres and Spherical Shells*, Air Force Cambridge Research Center, Geophysical Research Paper No. 15.

Aden, A. L., and Kerker, M., 1951, *J. Appl. Phys.*, **22**, 1242.

Anders, T., 1953, *Z. Phys.*, **135**, 219.

Anderson, D. V., 1952, *J. Acoust. Soc. Amer.*, **24**, 324.

Anderson, D. V., Northwood, T. D., and Barnes, C., 1952, *J. Acoust. Soc. Amer.*, **24**, 276.

Anderson, V. C., 1950, *J. Acoust. Soc. Amer.*, **22**, 426.

Andrejewski, W., 1951, *Naturwissenschaften*, **38**, 406; 1952, *Die Beugung elektromagnetischer Wellen an der leitenden Kreisscheibe und an der kreisförmigen Öffnung im leitenden ebenen Schirm, Dissertation*, University of Aachen; 1953, *Z. angew. Phys.*, **5**, 178.

Andrews, C. L., 1946, *Phys. Rev.*, **69**, 684 ; 1947, *Ibid.*, **71**, 777 ; 1948, *Ibid.*, **74**, 1231 ; 1950 a, *Ibid.*, **78**, 344 ; 1950 b, *J. Appl. Phys.*, **21**, 761, **22**, 520 ; 1951 a, *Amer. J. Phys.*, **19**, 280 ; 1951 b, *J. Appl. Phys.*, **22**, 465.

Aoki, T., 1949 a, *J. Phys. Soc. Japan*, **4**, 183 ; 1949 b, *Ibid.*, **4**, 186.

Aoki, T., Sunouchi, G., and Utagawa, M., 1949, *Sci. Rep. Tôhoku Univ.*, **33**, 38.

Artmann, K., 1950, *Z. Phys.*, **127**, 468.

Ashour, A. A., 1950, *Quart. J. Mech.*, **3**, 119 ; 1952, *Ibid.*, **5**, 379.

Atkinson, F. V., 1949, *Phil. Mag.*, **40**, 645.

Atlas, D., Kerker, M., and Hitschfeld, W., 1953, *J. Atmos. Terr. Phys.*, **3**, 108.

Awatani, J., 1951, *Mem. Inst. Sci. Res. Osaka Univ.*, **2**, 14.

Backhaus, H., 1943, *Z. tech. Phys.*, **24**, 75.

Bailin, L. L., 1951, *J. Res., Nat. Bur. Stand.*, **47**, 315.

Baker, B. B., and Copson, E. T., 1950, *The Mathematical Theory of Huygens' Principle* (Oxford : Clarendon Press).

Bardeen, J., 1930, *Phys. Rev.*, **36**, 1482.

Barnes, C., and Anderson, D. V., 1952, *J. Acoust. Soc. Amer.*, **24**, 229.

Barrow, W. L., and Chu, L. J., 1939, *Proc. Inst. Radio Engrs, N.Y.*, **27**, 51.

Bazer, J., and Karp, S. N., 1952, *Propagation of Plane Electromagnetic Waves past a Shoreline*, New York University, Mathematics Research Group, Research Report No. EM-46.

Begovich, N. A., 1950, *Proc. Inst. Radio Engrs, N.Y.*, **38**, 803.

Bekefi, G., 1952 a, *J. Appl. Phys.*, **23**, 1403; 1952 b, *Diffraction of Electromagnetic Waves by an Aperture in an Infinite Screen*, McGill University, The Eaton Electronics Research Laboratory, Technical Report No. 22 ; 1952 c, *Diffraction of Sound Waves by a Circular Aperture, Ibid.*, No. 16 ; 1953, *J. Acoust. Soc. Amer.*, **25**, 205.

Bekefi, G., and Woonton, G. A., 1952, *Microwave Diffraction Measurements on Circular Apertures*, McGill University, The Eaton Electronics Research Laboratory, Technical Report No. 24.

Berry, F. J., 1952 a, *Quart. J. Mech.*, **5**, 324 ; 1952 b, *Ibid.*, **5**, 333.

Berz, F., 1951, *Proc. Instn. Elect. Engrs*, Pt. III, **98**, 47.

Bethe, H. A., 1943 a, *Lumped Constants for Small Holes*, Radiation Laboratory Report, No. 194 ; 1943 b, *Theory of Side Windows in Waveguides, Ibid.*, No. 199 ; 1944, *Phys. Rev.*, **66**, 163.

Blass, J., 1951, *The Extension of the Wiener–Hopf Technique to Radiation Problems involving Boundaries of Elliptical Cross-Section, Dissertation*, Polytechnic Institute of Brooklyn.

Boersch, H., 1951, *Z. Phys.*, **131**, 78.

Booker, H. G., 1946, *J. Instn. Elect. Engrs*, Pt. III A, **93**, 620.

Booker, H. G., and Clemmov, P. C., 1950 a, *Proc. Instn Elect. Engrs*, Pt. III, **97**, 11 ; 1950 b, *Ibid.*, Pt. III, **97**, 18.

Booker, H. G., Ratcliffe, J. A., and Shinn, D. H., 1950, *Phil. Trans. Roy. Soc.* A, **242**, 579.

Born, H., 1943, *Hochfrequenztech. u. Elektroakust.*, **62**, 20.

Borts, R. B., and Woonton, G. A., 1952, *The Effect of the Directivity of the Probe upon the Measurement of Aperture Fields*, McGill University, The Eaton Electronics Research Laboratory, Technical Report No. 21.

BOURGIN, D. G., 1945, *Math. Rev.*, **6**, 165.

BOUWKAMP, C. J., 1940, *Physica, 's-Grav.*, **7**, 485 ; 1941, *Theoretische en numerieke behandeling van de buiging door een ronde opening, Dissertation*, University of Groningen ; 1946 a, *Philips Res. Rep.*, **1**, 251 ; 1946 b, *Physica, 's-Grav.*, **12**, 467 ; 1947 a, *Math. Rev.*, **8**, 179 ; 1947 b, *J. Math. Phys.*, **26**, 79 ; 1948, *Proc. Acad. Sci. Amst.*, **51**, 891 (=*Indagationes Math.*, **10**, 424) ; 1949 a, *Phys. Rev.*, **75**, 1608 ; 1949 b, *Proc. Acad. Sci. Amst.*, **52**, 987 (=*Indagationes Math.*, **11**, 366) ; 1950 a, *Physica, 's-Grav.*, **16**, 1 ; 1950 b, *Philips Res. Rep.*, **5**, 87 ; 1950 c, *Proc. Acad. Sci. Amst.*, **53**, 654 (=*Indagationes Math.*, **12**, 208) ; 1950 d, *Proc. Acad. Sci. Amst.*, **53**, 931 (=*Indagationes Math.*, **12**, 326); 1950 e, *Philips Res. Rep.*, **5**, 321 ; 1950 f, *Ibid.*, **5**, 401 ; 1953, *Diffraction Theory, a Critique of some Recent Developments*, New York University, Mathematics Research Group, Research Report No. EM–50.

BRAUMANN, H., 1948, *Z. Naturforsch.*, **3a**, 340.

BRAUNBEK, W., 1949, *Ann. Phys., Lpz.*, **6**, 53 ; 1950 a, *Z. Phys.*, **127**, 381 ; 1950 b, *Ibid.*, **127**, 405.

BRAUNBEK, W., and LAUKIEN, G., 1952, *Optik, Stuttgart*, **9**, 174.

BREMMER, H., 1949, *Terrestrial Radio Waves* (Amsterdam : Elsevier) ; 1951 a, *Commun. Pure Appl. Math.*, **4**, 61 ; 1951 b, *Physica, 's-Grav.*, **17**, 63.

BRILLOUIN, J., 1949, *C.R. Acad. Sci., Paris*, **229**, 513.

BRILLOUIN, L., 1949, *J. Appl. Phys.*, **20**, 1110.

BRODIN, J., 1949 a, *C.R. Acad. Sci., Paris*, **229**, 989 ; 1949 b, *Ibid.*, **229**, 1064 ; 1950 a, *Ibid.*, **230**, 1345 ; 1950 b, *Ibid.*, **230**, 1388.

DE BROGLIE, L., 1948, *Les ondes électromagnétiques centimétriques* (Paris : Revue d'Optique théor. et instr.) ; 1951, *Problèmes de propagations guidées des ondes électromagnétiques* (Paris : Gauthiers–Villars).

BUCERIUS, H., 1946, *Optik, Stuttgart*, **1**, 188.

BUCHHOLZ, H., 1942, *Ann. Phys., Lpz.*, **42**, 423 ; 1943 a, *Z. angew. Math. Mech.*, **23**, 47, 101 ; 1943 b, *Arch. Elektrotech.*, **37**, 22, 87, 145 ; 1948 a, *Ann. Phys., Lpz.*, **2**, 185 ; 1948 b, *Z. Phys.*, **124**, 196 ; 1948 c, *Arch. elektr. Übertr.*, **2**, 14 ; 1953, *Die konfluente hypergeometrische Funktion*, Ergebnisse der angewandten Mathematik, vol. 2 (Berlin : Springer Verlag).

BURROWS, C. R., and ATTWOOD, S. S., 1949, *Radio Wave Propagation, Technical Report National Defense Research Committee* (New York : Academic Press).

CAPRIOLI, L., 1949, *R. C. Accad. Lincei*, **7**, 107.

CARLSON, J. F., and HEINS, A. E., 1947, *Quart. Appl. Math.*, **4**, 313.

CARTER, A. H., and WILLIAMS, A. O., 1951, *J. Acoust. Soc. Amer.*, **23**, 179 ; 1952, *Ibid.*, **24**, 230.

CARTER, P. S., 1943, *Proc. Inst. Radio Engrs, N.Y.*, **31**, 671.

CASIMIR, H. B. G., 1951, *Philips Res. Rep.*, **6**, 162.

CHENEY, M. G., and WATSON, R. B., 1951, *J. Appl. Phys.*, **22**, 675.

CHESTER, W., 1950 a, *Phil. Mag.*, **41**, 11 ; 1950 b, *Phil. Trans. Roy. Soc.* A, **242**, 527 ; 1950 c, *Proc. Roy. Soc.* A, **203**, 33 ; 1952, *Quart. J. Mech.*, **5**, 196.

CHETAEV, D. N., 1951 a, *C.R. Acad. Sci. U.R.S.S.*, **76**, 813 ; 1951 b, *Appl. Math. Mech., Leningr.*, **15**, 439.

CHU, L. J., 1940, *J. Appl. Phys.*, **11**, 603.

CHU, L. J., and HANSEN, W. W., 1949, *J. Appl. Phys.*, **20**, 280.

CLAVIER, A. G., 1948, *Elect. Commun.*, **25**, 148.

CLEMMOW, P. C., 1950 a, *Quart. J. Mech.*, **3**, 241 ; 1950 b, *Ibid.*, **3**, 377 ; 1951, *Proc. Roy. Soc.* A, **205**, 286 ; 1953, *Phil. Trans. Roy. Soc.* A, **246**, 1.

COHN, S. B., 1951, *J. Appl. Phys.*, **22**, 628 ; 1952, *Proc. Inst. Radio Engrs, N.Y.*, **40**, 1069.

COMPUTATION LABORATORY NATIONAL BUREAU OF STANDARDS, 1951, *Tables relating to Mathieu Functions* (New York : Columbia University Press).

COOK, R. K., and CHRZANOWSKI, P., 1946, *J. Acoust. Soc. Amer.*, **17**, 315.

COPSON, E. T., 1946 a, *Quart. J. Math.*, **17**, 19 ; 1946 b, *Proc. Roy. Soc.* A, **186**, 100 ; 1950, *Ibid.*, A, **202**, 277 ; 1951, *Math. Rev.*, **12**, 774.

CROZE, F., and BOILLET, P., 1949, *C.R. Acad. Sci., Paris*, **228**, 305.

CROZE, F., and DARMOIS, G., 1949, *C.R. Acad. Sci., Paris*, **228**, 824.

CROZE, F., and DURAND, E., 1949, *C.R. Acad. Sci., Paris*, **228**, 236.

DA SILVEIRA, A., 1951, *C.R. Acad. Sci., Paris*, **233**, 1269.

DE POSSEL, R., and POUGET-MICHEL, C., 1951, *C.R. Acad. Sci., Paris*, **232**, 1819.

Di Jorio, M., 1950 a, *Ottica*, **4**, 58 ; 1950 b, *Ibid.*, **4**, 97 ; 1951, *Ibid.*, **5**, 9.

Dockès, J., 1952, *Câbles & Transm.*, **6**, 221.

Domb, C., 1953, *Advances in Physics* (*Phil. Mag.* Suppl.), **2**, 96.

Dörr, J., 1952, *Z. angew. Math. Phys.*, **3**, 427.

Dunning, K. L., and Fellers, R. G., 1951, *J. Appl. Phys.*, **22**, 1316.

Durand, E., 1948 a, *C.R. Acad. Sci., Paris*, **226**, 1440 ; 1948 b, *Ibid.*, **226**, 1593 ; 1948 c, *Ibid.*, **226**, 1812 ; 1948 d, *Ibid.*, **226**, 1972 ; 1949, *Rev. Opt.* (*théor. instrum.*), **28**, 325.

Eberlein, W. F., 1948, *Phys. Rev.*, **74**, 190.

Ehrlich, M. J., 1951, *Diffraction of a Plane Electromagnetic Wave by a Circular Aperture and Complementary Obstacle, Dissertation*, University of California.

Elliott, R. S., 1952, *Proc. Inst. Radio Engrs, N.Y.*, **40**, 729.

Elrod, H., 1952, *J. Acoust. Soc. Amer.*, **24**, 325.

Erler, K., 1940, *Hochfrequenztech. u. Elektroakust.*, **56**, 104.

Estrin, G., 1950, *J. Appl. Phys.*, **21**, 667.

Faran, J. J., 1951, *J. Acoust. Soc. Amer.*, **23**, 405 ; 1953, *Ibid.*, **25**, 155.

Faust, R. C., 1951, *Proc. Phys. Soc.* B, **64**, 105.

Feenberg, E., 1932, *Phys. Rev.*, **40**, 40.

Fel'd, Ya. N., 1948, *C.R. Acad. Sci. U.R.S.S.*, **60**, 1165.

Felsen, L. B., 1952, *Diffraction by Wedges and Corners, Dissertation*, Polytechnic Institute of Brooklyn.

Feuer, P., and Akeley, E. S., 1948, *J. Appl. Phys.*, **19**, 39.

Flammer, C., 1951, *Prolate Spheroidal Wave Functions*, Stanford Research Institute, Technical
Report No. 16 ; 1952, *The Vector Wave Solution of the Diffraction of Electromagnetic Waves by Circular Disks and Apertures, Ibid.*, No. 27.

Fock, V. A., 1942, *C.R. Acad. Sci. U.R.S.S.*, **36**, 133 ; 1944, *Rec. Math., Moscou*, **14**, 3 ; 1946 a, *J. Phys., Moscow*, **10**, 130 ; 1946 b, *Ibid.*, **10**, 399 ; 1948, *Phil. Mag.*, **39**, 149 ; 1950, *J. Exp. Theor. Phys.*, **20**, 961 ; 1951, *Adv. Phys. Sci., Moscow*, **43**, 587.

Foldy, L.-L., and Primakoff, H., 1945, *J. Acoust. Soc. Amer.*, **17**, 109.

Fox, E. N., 1948, *Phil. Trans. Roy. Soc.* A, **241**, 71 ; 1949, *Ibid.*, A, **242**, 6 ; 1952, *Proc. Roy. Soc.* A, **211**, 398.

Fox, F. E., 1940, *J. Acoust. Soc. Amer.*, **12**, 147.

Frahn, W., 1951, *Die Anwendung des Braunbek'schen Näherungsverfahrens auf die Beugung elektromagnetischer Wellen an der kreisförmigen Öffnung im vollkommen leitenden ebenen Schirm und an der vollkommen leitenden Kreisscheibe, Thesis*, University of Aachen.

Franz, W., 1948, *Z. Naturforsch.*, **3a**, 500 ; 1949, *Z. Phys.*, **125**, 563 ; 1950 a, *Ibid.*, **128**, 432 ; 1950 b, *Proc. Phys. Soc.* A, **63**, 925 ; 1952, *Z. angew. Math. Mech.*, **32**, 26.

Franz, W., and Deppermann, K., 1952, *Ann. Phys., Lpz.*, **10**, 361.

Freudenthal, H., 1939, *Compos. Math.*, **6**, 221.

Friedlander, F. G., 1946, *Proc. Roy. Soc.* A, **186**, 322 ; 1951, *Quart. J. Mech.*, **4**, 344.

Friedman, B., 1951 a, *The Dipole Field in an Inhomogeneous Atmosphere*, New York University, Mathematics Research Group, Research Report No. EM–28 ; 1951 b, *Commun. Pure Appl. Math.*, **4**, 317.

Friedman, B., and Russek, J., 1952, *Addition Theorems for Spherical Waves*, New York University, Mathematics Research Group, Research Report No. EM–44.

Garnir, H. G., 1952 a, *Bull. Soc. Roy. Sci. Liège*, p. 119 ; 1952 b, *Ibid.*, p. 328.

Ginzton, E. L., and Goodwin, P. S., 1950, *Proc. Inst. Radio Engrs, N.Y.*, **38**, 305.

Goudet, G., 1947, *Onde élect.*, **27**, 313.

Greenstadt, J., 1948, *The Calculation of Scattering by Thin Shells*, New York University, Mathematics Research Group, Research Report No. 172–10.

Groschwitz, E., and Hönl, H., 1952, *Z. Phys.*, **131**, 305.

Grosjean, C. C., 1952, *Nuovo Cim.*, **9**, 220.

Guptill, E. W., 1953, *Canad. J. Phys.*, **31**, 393.

Guptill, E. W., and MacDonald, A. D., 1952, *Canad. J. Phys.*, **30**, 119.

Gumprecht, R. O., and Sliepcevich, C. M., 1951 a, *Tables of Functions of First and Second Partial Derivatives of Legendre Polynomials* ; 1951 b, *Tables of Riccati Bessel Functions for Large Arguments and Orders* ; 1951 c, *Tables of Scattering Functions for Spherical Particles* (University of Michigan : Ann Arbor).

GUTIN, L., 1937, *Tech. Phys., U.S.S.R.*, **4**, 404.
GÜTTLER, A., 1952, *Ann. Phys., Lpz.*, **11**, 65.
GUTTON, H., 1946, *Onde élect.*, **26**, 459.
HANSON, E. T., 1933, *Phil. Trans. Roy. Soc.* A, **232**, 223.
HARDEN, B. N., 1952, *Proc. Instn Elect. Engrs*, Pt. III, **99**, 229.
HARRINGTON, R. F., 1953, *J. Appl. Phys.*, **24**, 547.
HART, R. W., 1951, *J. Acoust. Soc. Amer.*, **23**, 323.
HART, R. W., and MONTROLL, E. W., 1951, *J. Appl. Phys.*, **22**, 376.
HARTIG, E. O., 1950, *Circular Apertures and their Effects on Half-Dipole Impedances*, Cruft Laboratory, Technical Report No. 107.
HARTNAGEL, W., and KAPPLER, E., 1949, *Z. Naturforsch.*, **4a**, 498.
HAYCOCK, O. C., and WILEY, F. L., 1952, *Proc. Inst. Radio Engrs, N.Y.*, **40**, 349.
HEINS, A. E., 1948 a, *Quart. Appl. Math.*, **6**, 157 ; 1948 b, *Ibid.*, **6**, 215 ; 1948 c, *Amer. J. Math.*, **70**, 730 ; 1950 a, *Canad. J. Math.*, **2**, 210 ; 1950 b, *Quart. Appl. Math.*, **8**, 281 ; 1951, *J. Math. Phys.*, **30**, 164.
HEINS, A. E., and CARLSON, J. F., 1947, *Quart. Appl. Math.*, **5**, 82.
HEINS, A. E., and FESHBACH, H., 1947, *J. Math. Phys.*, **26**, 143.
HERGLOTZ, G., 1952, *Math. Ann.*, **124**, 219.
HERZFELD, K. F., 1911, *S.B. Akad. Wiss. Wien*, Abt. IIa, **120**, 1587.
HOGG, D. C., 1953, *J. Appl. Phys.*, **24**, 110.
HOLL, H., 1946, *Optik, Stuttgart*, **1**, 213 ; 1948, *Ibid.*, **4**, 173.
HONERJÄGER, R., 1949, *Ann. Phys., Lpz.*, **4**, 25.
HÖNL, H., 1952, *Z. Phys.*, **131**, 290.
HÖNL, H., and MAUE, A. W., 1952, *Z. Phys.*, **132**, 569.
HÖNL, H., and ZIMMER, E., 1953, *Z. Phys.*, **135**, 196.
HORTON, C. W., 1949 a, *Phys. Rev.*, **75**, 1263 ; 1949 b, *Proc. Inst. Radio Engrs, N.Y.*, **37**, 744 ; 1953, *J. Acoust. Soc. Amer.*, **25**, 188.
HORTON, C. W., and KARAL, F. C., 1950, *J. Acoust. Soc. Amer.*, **22**, 855; 1951, *J. Appl. Phys.*, **22**, 575.
HORTON, C. W., and WATSON, R. B., 1950, *J. Appl. Phys.*, **21**, 16.
HOUGHTON, H. G., and CHALKER, W. R., 1949, *J. Opt. Soc. Amer.*, **39**, 955.
HOUSTON, R. E., and NOBLE, R. H., 1951, *J. Appl. Phys.*, **22**, 1295; 1952, *Phys. Rev.*, **85**, 732.
HUANG, C., 1953 a, *On the Diffraction of Electromagnetic Waves by Annular, Elliptical and Rectangular Apertures*, Cruft Laboratory, Technical Report No. 163 ; 1953 b, Cruft Laboratory, Technical Report No. 164.
HUANG, C., and KODIS, R. D., 1951, *Diffraction by Spheres and Edges at 1·25 centimeters*, Cruft Laboratory, Technical Report No. 138.
HUBBARD, J. C., ZARTMAN, I. F., and LARKIN, C. R., 1947, *J. Opt. Soc. Amer.*, **37**, 832.
IIJIMA, T., 1952, *On the Electromagnetic Fields in case of Existence of a Semi-Infinite Hollow Conductive Circular Cylinder, II*, Researches of the Electrotechnical Laboratory, No. 531, Tokyo.
INGERSLEV, F., and NIELSEN, A. K., 1944, *Ingen. Vidensk. Skr.*, No. 5.
JAECKEL, R., 1951, *J. reine angew. Math.*, **189**, 141.
JOBST, G., 1925, *Ann. Phys., Lpz.*, **76**, 863.
JONES, D. S., 1950, *Quart. J. Mech.*, **3**, 420 ; 1952 a, *Ibid.*, **5**, 363 ; 1952 b, *Proc. Lond. Math. Soc.*, **2**, 440; 1952 c, *Quart. J. Math.*, **3**, 189; 1952 d, *Proc. Camb. Phil. Soc.*, **48**, 118 ; 1952 e, *Ibid.*, **48**, 733 ; 1953, *Proc. Roy. Soc.* A, **217**, 153.
JONES, D. S., and PIDDUCK, F. B., 1950, *Quart. J. Math.*, **1**, 229.
JUNGER, M. C., 1952, *J. Acoust. Soc. Amer.*, **24**, 366 ; 1953, *Ibid.*, **25**, 40.
KARP, S. N., 1950, *Commun. Pure Appl. Math.*, **3**, 411.
KARP, S., and SOLLFREY, W., 1950, *Diffraction by a Dielectric Wedge with Application to Propagation through a Cold Front*, New York University, Mathematics Research Group, Research Report No. EM–23.
KATHAVATE, Y. V., 1945 a, *Proc. Indian Acad. Sci.* A, **21**, 177 ; 1945 b, *Ibid.*, A, **21**, 233.
KAWAI, N., 1952, *Sci. Rep. Tôhoku Univ.*, **35**, 210.
KAY, I., 1952, *Diffraction of an Arbitrary Pulse by a Wedge*, New York University, Mathematics Research Group, Research Report No. EM–43.
KELLER, H. B., and KELLER, J. B., 1949, *J. Appl. Phys.*, **20**, 393.
KELLER, J. B., 1950, *J. Appl. Phys.*, **21**, 896 ; 1952, *Ibid.*, **23**, 1267.

KELLER, J. B., and BLANK, A., 1951, *Commun. Pure Appl. Math.*, **4**, 75.
KELLER, J. B., and KELLER, H. B., 1950, *J. Opt. Soc. Amer.*, **40**, 48.
KELLER, J. B., and PREISER, S., 1950, *Determination of Reflected and Transmitted Fields by Geometrical Optics, II*, New York University, Mathematics Research Group, Research Report No. EM–20.
KERKER, M., LANGLEBEN, P., and GUNN, K. L. S., 1951, *J. Meteorol.*, **8**, 424.
KING, L. V., 1934, *Canad. J. Res.*, **11**, 135 ; 1935, *Proc. Roy. Soc.* A, **153**, 1, 17.
KISUN'KO, G. V., 1949, *C.R. Acad. Sci. U.R.S.S.*, **66**, 863.
KLEIN, W., 1951, *Z. angew. Phys.*, **3**, 253.
KLOEPFER, W., 1952, *Theorie der Kreisscheibenantennen, Dissertation*, University of Aachen.
KOCK, W. E., and HARVEY, F. K., 1951, *Bell Syst. Tech. J.*, **30**, 564.
KODIS, R. D., 1950, *An Experimental Investigation of Microwave Diffraction*, Cruft Laboratory, Technical Report No. 105 ; 1952, *J. Appl. Phys.*, **23**, 249.
KONTOROWICH, M. J., and LEBEDEV, N. N., 1939, *J. Phys., Moscow*, **1**, 229.
KOTANI, M., 1933, *Proc. Phys.-Math. Soc. Japan*, **15**, 30.
KUPRADZE, V. D., 1935, *Fundamental Problems in the Mathematical Theory of Diffraction* (translation from the Russian by C. D. BENSTER, National Bureau of Standards, Rep. 2008, 1952) ; 1939, *Compos. Math.*, **6**, 228.
LABRUM, N. R., 1952, *J. Appl. Phys.*, **23**, 1320, 1324.
LAIRD, D. T., and COHEN, H., 1952, *J. Acoust. Soc. Amer.*, **24**, 46.
LAMAR, E. S., 1949, *J. Opt. Soc. Amer.*, **39**, 929.
LANDER, R. L., and NIELSEN, C. E., 1953, *Rev. Sci. Instrum.*, **24**, 20.
LAX, M., 1950, *Phys. Rev.*, **78**, 306.
LAX, M., and FESHBACH, H., 1948, *J. Acoust. Soc. Amer.*, **20**, 108.
LEBEDEV, N. N., 1949, *C.R. Acad. Sci. U.R.S.S.*, **65**, 621.
LEDINEGG, E., 1948, *S. B. öst. Akad. Wiss.*, Abt. IIa, **156**, 417.
LEDINEGG, E., and URBAN, P., 1952, *Ann. Phys., Lpz.*, **10**, 349.
LEITNER, A., 1949, *J. Acoust. Soc. Amer.*, **21**, 331 ; 1950, *Effect of a Circular Groundplane on Antenna Radiation*, New York University, Mathematics Research Group, Research Report, No. EM–19.
LEITNER, A., and SPENCE, R. D., 1950 a, *J. Franklin Inst.*, **249**, 299 ; 1950 b, *J. Appl. Phys.*, **21**, 1001.
LENGYEL, B. A., 1951, *J. Appl. Phys.*, **22**, 265.
LEONTOVIČ, M., 1946, *J. Exp. Theor. Phys.*, **16**, 474.
LEVIN, M. L., 1944, *J. Exp. Theor. Phys.*, **14**, 481 ; 1948 a, *J. Tech. Phys., Moscow*, **18**, 639 ; 1948 b, *Ibid.*, **18**, 653.
LEVINE, H., 1950, *J. Acoust. Soc. Amer.*, **22**, 48 ; 1951, *Ibid.*, **23**, 307.
LEVINE, H., and PAPAS, C. H., 1951, *J. Appl. Phys.*, **22**, 29.
LEVINE, H., and SCHWINGER, J., 1948 a, *Phys. Rev.*, **73**, 383 ; 1948 b, *Ibid.*, **74**, 958 ; 1948 c, *Ibid.*, **74**, 1212 ; 1949 a, *Ibid.*, **75**, 1423 ; 1949 b, *Ibid.*, **75**, 1608 ; 1950, *Commun. Pure Appl. Math.*, **3**, 355.
LEVITAS, A., and LAX, M., 1951, *J. Acoust. Soc. Amer.*, **23**, 316.
LEWIN, L., 1949, *Nature, Lond.*, **164**, 311 ; 1952, *Proc. Instn Elect. Engrs*, **99**, Pt. IV, 168.
LEWIS, E. A., and CASEY, J. P., 1952, *J. Appl. Phys.*, **23**, 605.
LJUNGGREN, T., 1948, *Ark. Mat. Astr. Fys.* A, **36**, No. 14 ; 1949, *Ark. Fys.*, **1**, No. 1.
LOWAN, A. N., MORSE, P. M., FESHBACH, H., and LAX, M., 1946, *Scattering and Radiation from Circular Cylinders and Spheres*, U. S. Navy Department, Office of Research and Inventions.
LUCKE, W. S., 1951 a, *J. Appl. Phys.*, **22**, 14 ; 1951 b, *Transmission through a Rectangular Aperture in an Infinite Screen*, Stanford Research Institute, Technical Report No. 25.
LUDWIG, C., 1951, *Zur Theorie der Kreisscheibenantennen, Thesis*, University of Aachen.
LUNEBERG, R. K., 1944, *Mathematical Theory of Optics* (Brown University).
MAGNUS, W., 1941 a, *Z. Phys.*, **117**, 168 ; 1941 b, *Ibid.*, **118**, 343 ; 1942, *Jber. dtsch. MatVer.*, **52**, 177 ; 1949, *Abh. math. Sem. Hamburg. Univ.*, **16**, 77 ; 1951, *Infinite Matrices associated with Diffraction by an Aperture*, New York University, Mathematics Research Group, Research Report No. EM–32 ; 1953, *Quart. Appl. Math.*, **11**, 77.
MAGNUS, W., and OBERHETTINGER, F., 1950, *Commun. Pure Appl. Math.*, **3**, 393.
MAGNUS, W., and RELLICH, F., 1943, *Jber. dtsch. MatVer.*, **53**, 57.

MARCUVITZ, N., 1949, *The Representation, Measurement, and Calculation of Equivalent Circuits for Waveguide Discontinuities with Application to Rectangular Slots*, Polytechnic Institute of Brooklyn ; 1951 a, *Waveguide Handbook* (New York : McGraw-Hill) ; 1951 b, *Commun. Pure Appl. Math.*, **4**, 263.

MARCUVITZ, N., and SCHWINGER, J., 1951, *J. Appl. Phys.*, **22**, 806.

MARÉCHAL, A., and NAUDIN, F., 1950, *C.R. Acad. Sci., Paris.*, **230**, 1515.

MAUE, A. W., 1949 a, *Z. Naturforsch.*, **4a**, 393 ; 1949 b, *Z. Phys.*, **126**, 601 ; 1952, *Z. Naturforsch.*, **7a**, 387.

MCDONALD, K. L., and HARRIS, F. S., 1952, *J. Opt. Soc. Amer.*, **112**, 321.

MCLACHLAN, N. W., 1947, *Theory and Applications of Mathieu Functions* (Oxford : Clarendon Press).

MEIXNER, J., 1946 a, *Nachr. Akad. Wiss. Göttingen*, **2**, 74 ; 1946 b, *Z. Naturforsch.*, **1**, 496 ; 1947, *Z. angew. Math. Mech.*, **25/27**, 137 ; 1948 a, *Ibid.*, **28**, 304 ; 1948 b, *Z. Naturforsch.*, **3a**, 506 ; 1949 a, *Lamé's Wave Functions of the Ellipsoid of Revolution* (Washington : NACA, Mem. 1224) ; 1949 b, *Ann. Phys., Lpz.*, **6**, 2 ; 1949 c, *Math. Nachr.*, **3**, 9 ; 1949 d, *Ibid.*, **3**, 14 ; 1949 e, *Arch. Math., Karlsruhe*, **1**, 432 ; 1950, *Math. Nachr.*, **3**, 193 ; 1951 a, *Ibid.*, **5**, 1 ; 1951 b, *Ibid.*, **5**, 371 ; 1953, *Ann. Phys., Lpz.*, **12**, 227.

MEIXNER, J., and ANDREJEWSKI, W., 1950, *Ann. Phys., Lpz.*, **7**, 157.

MEIXNER, J., and FRITZE, U., 1949, *Z. angew. Phys.*, **1**, 535.

MEIXNER, J., and KLOEPFER, W., 1951, *Z. angew. Phys.*, **3**, 171.

MEIXNER, J., and SCHÄFKE, F. W., 1954, *Mathieu'sche Funktionen und Spheroidfunktionen mit Anwendungen auf physikalische und technische Probleme* (Heidelberg : Springer).

MILES, J. W., 1946 a, *Proc. Inst. Radio Engrs, N.Y.*, **34**, 728 ; 1946 b, *J. Acoust. Soc. Amer.*, **17**, 259, 272 ; 1947 a, *Ibid.*, **19**, 572 ; 1947 b, *Ibid.*, **19**, 579 ; 1948 a, *Ibid.*, **20**, 370 ; 1948 b, *Phys. Rev.*, **74**, 1531 ; 1948 c, *J. Acoust. Soc. Amer.*, **20**, 652 ; 1949 a, *Quart. Appl. Math.*, **7**, 45 ; 1949 b, *J. Math. Phys.*, **28**, 223 ; 1949 c, *Phys. Rev.*, **75**, 695 ; 1949 d, *J. Acoust. Soc. Amer.*, **21**, 140, 434 ; 1949 e, *J. Appl. Phys.*, **20**, 760 ; **21**, 468 ; 1950, *J. Acoust. Soc. Amer.*, **22**, 59 ; 1952 a, *Acustica*, **2**, 287 ; 1952 b, *Proc. Roy. Soc.* A, **212**, 543 ; 1952 c, *Ibid.*, A, **212**, 547 ; 1952 d, *J. Acoust. Soc. Amer.*, **24**, 324.

MILLER, W., 1951, *J. Appl. Phys.*, **22**, 55.

MIRIMANOV, R. G., 1948 a, *C.R. Acad. Sci. U.R.S.S.*, **60**, 203 ; 1948 b, *Ibid.*, **60**, 357 ; 1948 c, *Ibid.*, **61**, 617 ; 1949, *Ibid.*, **67**, 65 ; 1950, *Ibid.*, **71**, 879.

MÖGLICH, F., 1927, *Ann. Phys., Lpz.*, **83**, 609.

MOLLOY, C. T., 1948, *J. Acoust. Soc. Amer.*, **20**, 387 ; 1950, *Ibid.*, **22**, 551.

MONTROLL, E. W., and GREENBERG, J. M., 1952, *Phys. Rev.*, **86**, 889.

MONTROLL, E. W., and HART, R. W., 1951, *J. Appl. Phys.*, **22**, 1278.

MORSE, P. M., and RUBENSTEIN, P. J., 1938, *Phys. Rev.*, **54**, 895.

MOTZ, H., 1951, *Electromagnetic Problems of Microwave Theory* (London : Methuen).

MOULLIN, E. B., and PHILLIPS, F. M., 1952, *Proc. Instn Elect. Engrs*, Pt. IV, **99**, 137.

MÜLLER, C., 1949 a, *Arch. Math., Karlsruhe*, **1**, 296 ; 1949 b, *Abh. math. Sem. hamburg. Univ.*, **16**, 95 ; 1950, *Abh. dtsch. Akad. Wiss. Berl.*, No. 3 (Math.-naturwiss. Kl. 1945/46, Berlin : Akademie Verlag) ; 1951, *Math. Ann.*, **123**, 345 ; 1952, *Math. Z.*, **56**, 80.

MULLER, G. G., BLACK, R., and DAVIS, T. E., 1938, *J. Acoust. Soc. Amer.*, **10**, 6.

MÜLLER, R., 1952, *Z. angew. Phys.*, **4**, 424 ; 1953 a, *Z. Naturforsch.*, **8a**, 56 ; 1953 b, *Z. angew. Phys.*, **5**, 142.

MÜLLER, R., and WESTPFAHL, K., 1953, *Z. Phys.*, **134**, 245.

NEUGEBAUER, H. E. J., 1952 a, *A Method for Solving Diffraction Problems by Approximation*, McGill University, The Eaton Electronics Research Laboratory, Technical Report No. 23 ; 1952 b, *J. Appl. Phys.*, **23**, 1406.

NIMURA, T., 1951, *Sci. Rep. Tôhoku Univ.* B, **1/2**, 381.

NIMURA, T., and SHIBAYAMA, K., 1951, *Sci. Rep. Tôhoku Univ.* B, **3**, 77.

NIMURA, T., and WATANABE, Y., 1953, *J. Acoust. Soc. Amer.*, **25**, 76.

NOMURA, Y., 1951, *Sci. Rep. Tôhoku Univ.* B, **1/2**, 25.

NOMURA, Y., and AIDA, Y., 1951, *Sci. Rep. Tôhoku Univ.* B, **1/2**, 337.

NOMURA, Y., and KAWAI, N., 1949 a, *Sci. Rep. Tôhoku Univ.*, **33**, 197 ; 1949 b, *Ibid.*, **33**, 208.

NOVOBATZKY, K. F., 1942, *Z. Phys.*, **119**, 102.

OBERHETTINGER, F., 1943, *Ann. Phys., Lpz.*, **43**, 136.
OLINER, A. A., 1951, *Equivalent Circuits for Slots in Rectangular Waveguides*, Polytechnic Institute of Brooklyn.
ORTUSI, J., 1945, *Ann. Radioélect.*, **1**, 87.
ORTUSI, J., and SIMON, J. C., 1950, *Ann. Radioélect.*, **5**, 321.
OSEEN, C. W., 1944, *Ark. Mat. Astr. Fys.* A, **31**, No. 4.
OSTER, G., 1948, *Chem. Rev.*, **43**, 319.
PACHNER, J., 1951, *J. Acoust. Soc. Amer.*, **23**, 185, 198, 481.
PALEY, R. E. A. C., and WIENER, N., 1934, *The Fourier Transform in the Complex Domain* (American Mathematical Society Colloquium Publication).
PALMER, L. S., 1952, *J. Appl. Phys.*, **23**, 289.
PAPAS, C. H., 1949, *J. Math. Phys.*, **28**, 227 ; 1950, *J. Appl. Phys.*, **21**, 318.
PAPAS, C. H., and KING, R., 1948, *J. Appl. Phys.*, **19**, 808; 1949, *Quart. Appl. Math.*, **7**, 175.
PEKERIS, C. L., 1944, *Phys. Rev.*, **66**, 351.
PELLAM, J. R., 1940, *J. Acoust. Soc. Amer.*, **11**, 396.
PERRIN, F., and ABRAGAM, A., 1951, *J. Phys. Radium*, **12**, 69.
PIDDUCK, F. B., 1946, *Phil. Mag.*, **37**, 280 ; 1947, *Ibid.*, **38**, 439 ; 1951, *Quart. J. Math.*, **2**, 316.
PINNEY, E., 1946, *J. Math. Phys.*, **25**, 49 ; 1947, *Ibid.*, **26**, 42.
PISTOLKORS, A. A., 1948, *Proc. Inst. Radio Engrs, N. Y.*, **36**, 56.
PRIMAKOFF, H., and KELLER, J. B., 1947, *J. Acoust. Soc. Amer.*, **19**, 820.
PRIMAKOFF, H., KLEIN, M. J., KELLER, J. B., and CARSTENSEN, E. L., 1947, *J. Acoust. Soc. Amer.*, **19**, 132, 515.
PRYCE, M. H. L., 1953, *Advances in Physics* (*Phil. Mag.* Suppl.), **2**, 67.
RAMACHANDRAN, G. N., 1945, *Proc. Indian Acad. Sci.* A, **21**, 165.
RAYLEIGH, Lord, 1897 a, *Phil. Mag.*, **43**, 259 (=1903, *Sci. Papers*, **4**, 283); 1897 b, *Ibid.*, **44**, 28 (=1903, *Ibid.*, **4**, 305) ; 1913, *Proc. Roy. Soc.* A, **89**, 194 (=1920, *Ibid.*, **6**, 375).
REESINCK, J. J. M., and DE VRIES, D. A., 1940, *Physica, 's-Grav.*, **7**, 603.
RELLICH, F., 1943, *Jber. dtsch. MatVer.*, **53**, 57.
RHODES, D. R., 1950, *J. Appl. Phys.*, **21**, 1181.
ROBINSON, H. L., 1953, *J. Appl. Phys.*, **24**, 35.
ROCARD, Y., 1946, *Onde élect.*, **26**, 288.
ROW, R. V., 1952, *Microwave Diffraction Measurements in a Parallel-Plate Region*, Cruft Laboratory, Technical Report No. 153.
RUCH, E., 1950, *Ann. Phys., Lpz.*, **7**, 248.
RUZE, J., and YOUNG, M., 1951, *J. Appl. Phys.*, **22**, 277.
RYDBECK, O. E. H., 1944, *Chalmers tek. Högsk. Handl.*, No. 34 ; 1945, *Ibid.*, No. 43.
RYDE, J. W., 1946, *Meteorological Factors in Radio Wave Propagation* (London: Physical Society), p. 169.
RŽEVKIN, S. N., 1951, *J. Tech. Phys., Moscow*, **21**, 1224.
SAITÔ, N., and IKEDA, Y., 1951, *J. Phys. Soc. Japan*, **6**, 305.
SAKURAI, T., 1950 a, *J. Phys. Soc. Japan*, **5**, 203 ; 1950 b, *Ibid.*, **5**, 389.
SASAO, M., 1932, *Proc. Phys.-Math. Soc. Japan*, **14**, 510.
SAUNDERS, W. K., 1950, *Uniqueness of Solution of the Exterior Problem of the Electromagnetic Field*, University of California, Antenna Laboratory, Issue No. 175 ; 1951, *Existence of the Solution of the Exterior Problem of the Electromagnetic Field*, *Ibid.*, No. 176 ; 1952, *Proc. Nat. Acad. Sci., Wash.*, **38**, 342.
SAVORNIN, J., 1939, *Ann. Phys., Paris*, **11**, 129.
SAVORNIN, J., and LAURENTJOYE, R., 1946, *Rev. Opt.* (*théor. instrum.*), **25**, 73.
SAXON, D. S., 1945, *Notes on Lectures by Julian Schwinger, Discontinuities in Waveguides*, Radiation Laboratory, Massachusetts Institute of Technology.
SCHACHENMEIER, R., 1951, *Arch. elektr. Übertr.*, **5**, 267.
SCHEFFERS, H., 1942, *Ann. Phys., Lpz.*, **42**, 211.
SCHELKUNOFF, S. A., 1936, *Bell Syst. Tech. J.*, **15**, 92 ; 1939, *Phys. Rev.*, **56**, 308 ; 1951, *Commun. Pure Appl. Math.*, **4**, 43.
SCHOCH, A., 1941, *Akust. Z.*, **6**, 318 ; 1950, *Ergebn. exakt. Naturwiss.*, **23**, 127.
SCHORR, M. G., and BECK, F. J., 1950, *J. Appl. Phys.*, **21**, 795.
SCHOUTEN, J. P., and DE HOOP, A. T., 1952, *Tijdschr. ned. Radiogenoot.*, **17**, 45.
SCHWARZ, L., 1943, *Akust. Z.*, **8**, 91.
SEILER, W., 1931, *Hochfrequenztech. u. Elektroakust.*, **37**, 79.

SENIOR, T. B. A., 1952, *Proc. Roy. Soc.* A, **213**, 436 ; 1953, *Quart. J. Mech.*, **6**, 101.
SEVERIN, H., 1946, *Z. Naturforsch.*, **1**, 487 ; 1950 a, *Z. Phys.*, **128**, 108 ; 1950 b, *Z. angew. Phys.*, **2**, 499 ; 1951, *Z. Phys.*, **129**, 426.
SEVERIN, H., and STARKE, C., 1952, *Acustica–Akust. Beih.*, **2**, 59.
SEVERIN, H., and VON BAECKMANN, W., 1951, *Z. angew. Phys.*, **3**, 22.
SEVICK, J., and STORER, J. E., 1952, *A General Theory of Plane-Wave Scattering from Finite Conducting Obstacles with Application to the Two-Antenna Problem*, Cruft Laboratory, Technical Report, No. 149.
SHEINGOLD, L. S., 1953, *J. Appl. Phys.*, **24**, 415.
SHMOYS, J., 1951, *J. Opt. Soc. Amer.*, **41**, 324 ; 1953, in course of publication New York University, Mathematics Research Group.
ŠÍFRIN, K. S., 1950 a, *Bull. Acad. Sci. U.R.S.S.* (Sér. Geograph. Geophys.), **14**, 64 ; 1950 b, *Ibid.*, **14**, 128.
SILVER, S., 1949, *On the Integral Equations of Planar Diffraction Problems*, University of California, Antenna Laboratory, Report No. 163.
SILVER, S., and EHRLICH, M. J., 1951, *Diffraction of a Plane Electromagnetic Wave by a Circular Aperture and Complementary Obstacle: Part I, High Frequency Approximation*, University of California, Antenna Laboratory, Issue No. 181.
SILVER, S., EHRLICH, M. J., and HELD, G., 1952, *Diffraction of a Plane Electromagnetic Wave by a Circular Aperture and Complementary Obstacle: Part II, Discussion of Experimental Results*, University of California, Antenna Laboratory, Issue No. 185.
SILVER, S., and SAUNDERS, W. K., 1950 a, *J. Appl. Phys.*, **21**, 153 ; 1950 b, *Ibid.*, **21**, 745.
SIMON, J. C., 1951, *Ann. Radioélect.*, **6**, 205.
SINCLAIR, D., 1947, *J. Opt. Soc. Amer.*, **37**, 475.
SINCLAIR, G., 1948, *Proc. Inst. Radio Engrs., N.Y.*, **36**, 1487 ; 1951, *Ibid.*, **39**, 660.
SIVIAN, L. J., and O'NEIL, H. T., 1932, *J. Acoust. Soc. Amer.*, **3**, 483.
SKAVLEM, S., 1951, *Arch. Math. Naturvid.*, **51**, 61.
SLETTEN, C. J., 1952, *Electromagnetic Scattering from Wedges and Cones*, Air Force Cambridge Research Center, Antenna Laboratory.
SMYTHE, W. R., 1947, *Phys. Rev.*, **72**, 1066.
SOLLFREY, W., 1949, *The Variational Solution of Scattering Problems*, New York University, Mathematics Research Group, Research Report No. EM–11 ; 1952, *Diffraction of Pulses by Conducting Wedges and Cones*, *Ibid.*, No. EM–45.
SOLLFREY, W., and SHMOYS, J., 1952, *Diffraction of Electromagnetic Waves by a Plane Wire Grating II*, New York University, Mathematics Research Group, Research Report No. EM–41.
SOMMERFELD, A., 1942, *Ann. Phys., Lpz.*, **42**, 389 ; 1948 a, *Ibid.*, **2**, 85 ; 1948 b, *Vorlesungen über theoretische Physik*, Vol. 6, 2nd Edn (Leipzig : Geest u. Portig), pp. 216–227, 282–291 ; 1950, *Ibid.*, Vol. 4, 1st Edn (Wiesbaden : Dieterich).
SPENCE, R. D., 1948 a, *J. acoust. Soc. Amer.*, **20**, 380 ; 1948 b, *Phys. Rev.*, **73**, 1267 ; 1949, *J. Acoust. Soc. Amer.*, **21**, 98.
SPENCE, R. D., and GRANGER, S., 1951, *J. Acoust. Soc. Amer.*, **23**, 701.
SPENCE, R. D., and LEITNER, A., 1948, *Phys. Rev.*, **74**, 349.
SPENCE, R. D., and WELLS, C. P., 1951, *Commun. Pure Appl. Math.*, **4**, 95.
SPENCER, D. E., 1951, *J. Appl. Phys.*, **22**, 386.
STENZEL, H., 1942, *Ann. Phys., Lpz.*, **41**, 245 ; 1949, *Ibid.*, **4**, 303 ; 1952 a, *Acustica*, **2**, 263 ; 1952 b, *J. Acoust. Soc. Amer.*, **24**, 417.
STEVENSON, A. F., 1948, *J. Appl. Phys.*, **19**, 24.
STORER, J. E., 1951, *J. Appl. Phys.*, **22**, 1058 ; 1952, *Ibid.*, **23**, 588.
STORRUSTE, A., 1948, *K. Norske Vid. Selsk. Forh.*, **21**, 84, 88.
STORRUSTE, A., and WERGELAND, H., 1948 a, *K. Norske Vid. Selsk. Forh.*, **21**, 38; 1948 b, *Phys. Rev.*, **73**, 1397.
STRATTON, J. A., and CHU, L. J., 1939 a, *Phys. Rev.*, **56**, 99 ; 1939 b, *Ibid.*, **56**, 316.
STRATTON, J. A., MORSE, P. M., CHU, L. J., and HUTNER, R. A., 1941, *Elliptic Cylinder and Spheroidal Wave Functions* (New York : Wiley; London : Chapman and Hall).
STRUTT, M. J. O., 1931, *Z. Phys.*, **69**, 597.
STUMPERS, F. L., 1953, *A Bibliography in Information Theory, Communication Theory, and Cybernetics*, Research Laboratory of Electronics, Massachusetts Institute of Technology.
SWENSON, G. W., and JOHNSON, W. E., 1952, *J. Acoust. Soc. Amer.*, **24**, 84.

Svešnikov, A. G., 1950, *C.R. Acad. Sci. U.R.S.S.*, **73,** 917.

Tai, C. T., 1948, *Reflection of Plane Electromagnetic Waves from a Perfectly Conducting Grounded Half-Cylinders*, Cruft Laboratory, Technical Report No. 40 ; 1952 a, *J. Appl. Phys.*, **23,** 909 ; 1952 b, *Quasi-Static Solution for Diffraction of a Plane Electromagnetic Wave by a Small Oblate Spheroid*, Stanford Research Institute, Technical Report No. 24 ; 1952 c, *On the Theory of Diffraction of Electromagnetic Waves by a Sphere, Ibid.*, No. 29 ; 1952 d, *Radiation from a Uniform Circular Loop Antenna in the Presence of a Sphere, Ibid.*, No. 32.

Takahashi, I., Watanabe, T., and Tanimoto, K., 1950, paper read before the Physical Society of Japan, 2nd November, 1950.

Teichmann, T., and Wigner, E. P., 1953, *J. Appl. Phys.*, **24,** 262.

Theimer, O., Wassermann, G. D., and Wolf, E., 1952, *Proc. Roy. Soc.* A, **212,** 426.

Theissing, H. H., 1950, *J. Opt. Soc. Amer.*, **40,** 232.

Titchmarsh, E. C., 1948, *Introduction to the Theory of Fourier Integrals* (Oxford : Clarendon Press).

Toraldo di Francia, G., 1948, *R. C. Accad. Lincei*, **4,** 730 ; 1949, *Rev. Opt.* (*Théor. Instrum.*), **28,** 597 ; 1950 a, *R.C. Accad. Lincei*, **8,** 369 ; 1950 b, *Nuovo Cim.*, **7,** 967.

Tranter, W. O., 1948, *Phys. Rev.*, **73,** 184.

Twersky, V., 1950, *J. Acoust. Soc. Amer.*, **22,** 539 ; 1951 a, *Ibid.*, **23,** 336 ; 1951 b, *J. Appl. Phys.*, **22,** 825 ; 1952 a, *Ibid.*, **23,** 407 ; 1952 b, *Ibid.*, **23,** 1099 ; 1952 c, *J. Acoust. Soc. Amer.*, **24,** 42.

Vaĭnšteĭn, L. A., 1947, *C.R. Acad Sci. U.R.S.S.*, **58,** 1957 ; 1948 a, *Bull. Acad. Sci. U.R.S.S.* (Sér. Phys.), **12,** 144 ; 1948 b, *Ibid.*, **12,** 166 ; 1948 c, *J. Tech. Phys., Moscow*, **18,** 1543 ; 1949, *Ibid.*, **19,** 911 ; 1950 a, *C.R. Acad. Sci., U.S.S.R.*, **74,** 485 ; 1950 b, *Ibid.*, **74,** 909 ; 1951 a, *J. Tech. Phys., Moscow*, **21,** 328 ; 1951 b, *Ibid.*, **21,** 346.

Van de Hulst, H. C., 1946, *Optics of Spherical Particles* (Amsterdam : Duwaer & Zonen); 1949 a, *Physica, 's-Grav.*, **15,** 740 ; 1949 b, *Rech. astr. Obs. Utrecht*, **11,** Pt. 2 ; 1950, *Astrophys. J.*, **112,** 1.

Van Vleck, J. H., Bloch, F., and Hamermesh, M., 1947, *J. Appl. Phys.*, **18,** 274.

Vasseur, J. P., 1948, *C.R. Acad. Sci., Paris*, **226,** 885 ; 1949 a, *Ibid.*, **229,** 179 ; 1949 b, *Ibid.*, **229,** 586 ; 1952 a, *Onde élect.*, **32,** 3 ; 1952 b, *Ibid.*, **32,** 55 ; 1952 c, *Ibid.*, **32,** 97 ; 1952 d, *Ann. Phys., Paris*, **7,** 506.

Watson, R. B., and Horton, C. W., 1950, *J. Appl. Phys.*, **21,** 802.

Wergeland, H., 1945, *Avh. Norsk. Vid. Akad., Oslo* (Mat.-Nat. Kl.), **1,** No. 9.

Weyl, H., 1919, *Ann. Phys., Lpz.*, **60,** 481 ; 1952 a, *Math. Z.*, **55,** 187 ; 1952 b, *Ibid.*, **56,** 105.

Wiegrefe, A., 1913, *Ann. Phys., Lpz.*, **42,** 1241.

Wiener, F. M., 1947, *J. Acoust. Soc. Amer.*, **19,** 444 ; 1948, *Ibid.*, **20,** 367 ; 1949 a, *Ibid.*, **21,** 39 ; 1949 b, *Ibid.*, **21,** 334 ; 1950, *Ibid.*, **22,** 47 ; 1951, *Ibid.*, **23,** 697.

Wolf, E., 1951, *Reports on Progress in Physics*, **14,** 95 (London : Physical Society).

Woodson, R. A., 1942, *Phys. Rev.*, **62,** 297.

Woonton, G. A., 1950 a, *Canad. J. Res.* A, **28,** 120 ; 1950 b, *J. Appl. Phys.*, **21,** 577 ; 1952 a, *Ibid.*, **23,** 1405 ; 1952 b, *The Probe Antenna and the Diffraction Field*, McGill University, The Eaton Electronics Research Laboratory, Technical Report No. 17.

Woonton, G. A., Carruthers, J. A., Elliott, H. A., and Rigby, E. C., 1951, *J. Appl. Phys.*, **22,** 390.

Zernov, N. V., 1951 a, *J. Tech. Phys., Moscow*, **21,** 766 ; 1951 b, *Ibid.*, **21,** 1066.

Zuhrt, H., 1947 a, *Frequenz*, **1,** 33 ; 1947 b, *Ibid.*, **1,** 63 ; 1948, *Ibid.*, **2,** 6.

Note. These references cover the period from 1940 up to May 1953.

ATMOSPHERIC ELECTRICITY

By J. ALAN CHALMERS
Department of Physics, Durham Colleges in the University of Durham

CONTENTS

PAGE

§ 1. Introduction 102
§ 2. Ionization in the atmosphere 102
2.1. Ions and kernels
2.2. Combination coefficients
2.3. Equilibrium conditions
2.4. Measurement of charged and uncharged kernels
2.5. Sizes of kernels
2.6. Absence of equilibrium
2.7. Variation of P
§ 3. Fine-weather fields and currents 105
3.1. Field current and conductivity
3.2. Measurement of field by collectors
3.3. Measurement by bound charge
3.4. The field mill
3.5. The agrimeter
3.6. Change of field with height
3.7. The idea of convection currents
3.8. Direct measurement of conduction current
3.9. Indirect measurement of conduction current
3.10. Variation of conductivity with height
3.11. Measurement of field on mountains
§ 4. Point discharge currents 109
4.1. The process of point discharge
4.2. Distinction between atmospheric and laboratory point discharge
4.3. Conditions of point discharge
4.4. Relation between point discharge current and field
4.5. Theoretical considerations
4.6. Space charge from point discharge
4.7. The alti-electrograph
4.8. Multiple points
4.9. Total point discharge current from cloud
4.10. Effective separation of discharging points
§ 5. Precipitation currents 113
5.1. Importance of precipitation currents
5.2. General results
5.3. Relation with field
5.4. Single drops
5.5. Interpretation of charge–field relation
5.6. Origin of rain charges
5.7. Rain charges in low fields
5.8. The ' mirror image effect '
5.9. Field and currents in fog and mist
§ 6. Maintenance of the earth's charge 117
6.1. Fine-weather current
6.2. Variation with time of day
6.3. Correlation with thunderstorms
6.4. Effect of induced charges
6.5. Currents below storms
6.6. Currents above thunderclouds
§ 7. Thunderclouds 120
7.1. Development of clouds
7.2. Methods of electrical investigation
7.3. Results of field and field change measurements
7.4. Currents in thunderclouds
7.5. Currents above and below clouds
7.6. Alti-electrograph results
7.7. Radiosonde results
7.8. Fields inside clouds
7.9. Charges on particles in clouds
7.10. Other results
7.11. Lightning results
7.12. The lower positive charge
7.13. The views of Wichmann
§ 8. The origin of cloud charges 125
8.1. The main charges in clouds
8.2. Subsidiary charges in clouds
8.3. Requirements of theory
8.4. Classification of theories
8.5. Wilson's theory
8.6. Wall's theory
8.7. Frenkel's theory
8.8. Ice impact theory
8.9. Findeisen's theory
8.10. Riming effects
8.11. Other theories
8.12. Discussion of theories
8.13. The lower positive charge
8.14. Subsidiary charges
8.15. Charge separation in nimbostratus clouds
§ 9. Conclusion 132
Acknowledgments 133
References 133

Abstract. A number of current problems in atmospheric electricity are discussed in turn; these are: ionization equilibrium in the atmosphere; fine weather fields and currents, particularly conditions close to the earth's surface; point discharge; precipitation currents and their origin; the maintenance of the charge on the earth; the electrical structure of thunderclouds; and theories of the origin of charges in clouds.

§1. Introduction

The subject of atmospheric electricity has not previously been dealt with in these Reports, except that portion of it covered by "Lightning" (Meek and Perry 1944–45). The author's book (Chalmers 1949) has given an account of the subject, including material to hand up to the end of 1947. In the present Report it is proposed to select a few of the problems in the subject in which there have been significant advances in recent work or in work (e.g. German war-time work) which has only become available since the book was written. In order to make each problem clear to those without much previous knowledge of the subject, in each case a brief survey of the subject is given.

Since the preparation of this Report was commenced a review by Wormell (1953) has appeared, covering some of the same subject matter. Although Wormell's review may be said to have 'stolen some of the thunder', or, perhaps more exactly, some of the thunderstorm, most of the present Report has been written as planned. The author feels that to have omitted some sections which have been covered by Wormell's review would have left this Report somewhat unbalanced and unrepresentative of the subject; where there is an overlap, a difference in viewpoint may, perhaps, add to the interest.

The author has not been able to take into consideration any work which had not come to his notice by the end of May 1953.

§2. Ionization in the Atmosphere

2.1. *Ions and Kernels*

Measurements have shown that there exist in the atmosphere 'small' ions with mobility (velocity in cm sec^{-1} in a field of 1 v cm^{-1}) of the order of 1 to 2 units, and 'large' ions with mobility of the order of 0·0003 unit. Some workers have also found evidence for the existence of 'intermediate' ions, but these are usually not taken into account in discussions of ionization equilibrium.

It appears to be fairly well confirmed that the initial process of ionization consists almost always in the removal of an electron from a molecule. The electron quickly attaches itself to another molecule and both molecules then gather round themselves other molecules to constitute small ions. The process by which large ions are formed is indirect and much slower.

There exist in the atmosphere particles which have become known as 'condensation nuclei' or 'Aitken nuclei'. With the use of the word 'nucleus' in a very different connection in atomic physics, it seems that the time has now arrived for a new name for the atmospheric particles, and the word 'kernel' is suggested and will be used here. (The German word for such a particle is 'Kern'.) The kernels are the centres on which condensation occurs when the air is saturated, and this provides the only means so far devised for counting kernels. It has been proved experimentally that the large ions are just kernels carrying charges, and are produced when a neutral kernel combines with a small

ion; a charged kernel can lose its charge by combining with an ion, large or small, of opposite sign.

There are processes by which large and small ions are produced and cease to exist, and it has been thought that a dynamic equilibrium can be set up with steady numbers of ions of both sizes and of both charges. It is this problem of equilibrium with which we shall deal in this section.

2.2. *Combination Coefficients*

If n_1, n_2 are the number of positive and negative small ions respectively, and N_1, N_2, N_0 the number of positive, negative and uncharged kernels respectively, in unit volume in each case, then it seems reasonable to assume that, for example, the rate at which positive kernels are produced by combination of small positive ions and uncharged kernels is proportional to n_1N_0, the factor of proportionality being called a 'combination coefficient'. There are a number of different combination coefficients which can be defined from the following equations:

$$\frac{dn_1}{dt} = q - \alpha n_1 n_2 - \eta_{12} n_1 N_2 - \eta_{10} n_1 N_0$$

$$\frac{dn_2}{dt} = q - \alpha n_1 n_2 - \eta_{21} n_2 N_1 - \eta_{20} n_2 N_0$$

$$\frac{dN_1}{dt} = \eta_{10} n_1 N_0 - \eta_{21} n_2 N_1 - \gamma N_1 N_2$$

$$\frac{dN_2}{dt} = \eta_{20} n_2 N_0 - \eta_{12} n_1 N_2 - \gamma N_1 N_2$$

where q denotes the number of pairs of small ions produced in unit volume in unit time.

The values of α and the η's are of the order of 10^{-5} to 10^{-6}, while γ is of the order of 10^{-9}, and can always be neglected.

2.3. *Equilibrium Conditions*

When there is equilibrium, then the variations with time are zero, and thus $\eta_{10} n_1 N_0 = \eta_{21} n_2 N_1$ and $\eta_{20} n_2 N_0 = \eta_{12} n_1 N_2$ as first deduced by Nolan and de Sachy (1927).

It is usually found that $N_1 = N_2 = N$ (say), so that

$$\frac{N}{N_0} = \left(\frac{\eta_{10}\eta_{20}}{\eta_{12}\eta_{21}}\right)^{1/2} \quad \text{and} \quad \frac{n_1}{n_2} = \left(\frac{\eta_{21}\eta_{20}}{\eta_{12}\eta_{10}}\right)^{1/2}.$$

Thus, if the η's remain the same, the value of the ratio N/N_0 should be a constant. Often it is found convenient to use $Z = N_1 + N_2 + N_0$, the total number of kernels, and the ratio $P = Z/(N_1 + N_2)$ or the ratio Z/N_0, which is $P/(P-1)$.

Measurements of these ratios have been made in a number of places, under varying conditions and these have been tabulated, with comments, by Israël (1941). A glance at this table shows that the expectation that P should be constant is very far from being fulfilled, and in many cases there is evidence for a variation of P with Z or with meteorological conditions, e.g. relative humidity.

2.4. *Measurement of Charged and Uncharged Kernels*

The measurement of the number of charged and uncharged kernels has been made by two methods. In the first method the number of kernels is counted directly, and after the removal of the charged kernels. In the second method

the number of charged kernels is found by an ion counter, which gives, in fact, the total charge carried by the ions; it must then be assumed that the charged kernels carry single charges, to obtain the number. A comparison of results by both methods by Nolan, Boylan and de Sachy (1925) led to the conclusion that kernels are never multiply charged, but this was contradicted by later work by Torreson and Wait (1934) and Hogg (1934).

If kernels can carry multiple charges, the simple equilibrium theory given above will not hold and it is necessary to introduce terms involving n_1N_1, etc. Bricard (1949) has given equations which bring in the multiply charged kernels.

A photoelectric counter has been developed by Nolan and Pollak (1946) and this has made the counting of kernels much easier, as well as allowing simultaneous measurements with two counters; in this the photoelectric cell measures the opacity produced by the condensation of water on the kernels.

2.5. *Sizes of Kernels*

Diffusion measurements of kernels by Nolan and Guerrini (1935) and by Nolan, Nolan and Gormley (1938) have made it possible to determine the radii of the kernels. Measurements with kernels derived from heated platinum and stored to give equilibrium gave a definite relation between the value of Z/N_0 and the radius (Nolan and Kennan 1949), which agrees with results deduced from formulae of Whipple (1933) and Harper (1934, 1935) except for large kernels, which may be multiply charged (Nolan 1950).

Measurements of Z/N_0 for the free atmosphere and simultaneous measurements of the size of kernels (Nolan and Doherty 1950) has shown that the value of Z/N_0 is much less than the equilibrium value to be expected theoretically.

2.6. *Absence of Equilibrium*

Calculations from the first formula of §2.2,

$$\frac{dn_1}{dt} = q - \alpha n_1 n_2 - \eta_{12} n_1 N_2 - \eta_{10} n_1 N_0,$$

had shown that the small ions reach their equilibrium conditions in a time of only a few seconds; it was then assumed that equilibrium of charged and uncharged kernels would be set up in a similar time.

But Israël (1948) first showed that, in fact, the half-value time for kernels is 5½ minutes, under the conditions existing, and Nolan (1950), independently, found a time of 10 to 15 minutes. Thus in most measurements of kernels in the free atmosphere it is unlikely that equilibrium has been reached, particularly at sites close to the place of origin of the kernels.

2.7. *Variation of P*

The failure of the simple result given in §2.3 for the ratio P may be ascribed to various factors, some of which have already been discussed. The constancy of P depends on the assumptions that kernels are not multiply charged, that equilibrium conditions exist, and that the η's are constant. We have already pointed out that the first two assumptions are incorrect, and it has been found that the sizes of kernels vary and depend, among other things, on humidity (Nolan and Tedde 1950), and clearly the η's depend on the kernel size.

Also, in the free atmosphere near the ground, the numbers of positive and negative ions vary with height, so that, in any volume of air, more ions enter the

volume by conduction than leave it, thus giving yet another correction to the simple equilibrium formula; this term has been considered by Chalmers (1946) in connection with another problem.

Although it has not yet been possible to give any general account of the values of P, it is easy to see why it is not constant, as would be expected from the simple theory.

§ 3. Fine-Weather Fields and Currents

3.1. *Field Current and Conductivity*

The most widely measured element of atmospheric electricity is the field, or potential gradient, F, near the earth's surface. The conductivity λ of the atmosphere near the surface has also been measured; the product of F and λ gives the density of the conduction current i, and it has become more and more evident, as the subject has progressed, that i is of more fundamental importance than either F or λ.

If the ionosphere, the conducting layer in the upper atmosphere, has a potential V with respect to the earth, and if a column of unit cross section from the earth to the ionosphere has a resistance R, then $i = V/R$.

Thus $F = i/\lambda = V/\lambda R$, and measurements of F depend on local variations of λ, and are often not of much significance. But where there are no such local variations, the field varies with V (see § 6.2).

Associated with a field F there must be a charge σ per unit area on the earth's surface where $F = -4\pi\sigma$, the negative sign arising from the usual convention in atmospheric electricity of a positive field being such that the potential increases on rising.

In the same way, if the field varies with height, there must be a space charge density ρ such that $dF/dh = -4\pi\rho$.

If there is a conduction current density which is the same at different heights, and if the conductivity varies with height, then the field F must vary with height, giving a space charge. Kasemir (1950) has pointed out that, while this space charge complicates matters if the fields and currents are looked upon as derived from the charges present, it is quite unnecessary to discuss the space charges at all if the current flowing is looked upon as the fundamental quantity.

3.2. *Measurement of Field by Collectors*

The older method of measurement of the vertical electric field in the atmosphere was to find the potential at a point 1 m (or some other convenient height) above the earth by the use of a 'potential equalizer', most usually a radioactive collector. With an electrometer the potential can be measured and recorded. Kasemir (1952) has used electronic methods to amplify the current produced so as to display it on a more robust instrument for a portable apparatus.

Mühleisen (1951) has shown that there are grave disadvantages in the use of a radioactive collector, particularly in regard to the effect of wind.

3.3. *Measurement by Bound Charge*

In the normal fine-weather field the lines of force in the atmosphere end on negative charges on the earth. If the negative charge on an area of the earth can be measured, this gives a method of measuring the field which is free from many of the objections of the collector method.

If an insulated plate in the plane of the earth is connected to a measuring instrument and is covered over with an earthed plate the bound negative charge is released and is measured. Wilson (1916) first used this method with an electroscope and with a capillary electrometer. Chalmers (1952a) has used the same method with a valve amplifier for transportable apparatus.

3.4. *The Field Mill*

If, instead of covering a plate once and measuring the charge released, the plate is alternately covered and uncovered in rapid succession and the measuring instrument is replaced by the grid of a valve, with a leak to earth, then the grid receives an alternating potential which can be amplified readily. The potential is proportional to the field, but does not give the sign of the field unless some other device is used. It seems that this method of measuring field was first used by Mathias (1926).

In order to obtain the sign of the field, various methods have been used, for example Waddel, Drutowski and Blatt (1946) rectified the output with a mechanical commutator driven off the same shaft as the covering device; Rangs (1942) drove an electromagnetic generator (dynamo) off the same motor as the covering device and added the output to that of the electrostatic generator, shorting out the latter at intervals to give a zero reading; Mapleson (1953) and Whitlock (1953) used electronic phase-sensitive rectification, using a generator on the same shaft as the covering device to provide the reference voltage, Mapleson using a dynamo and Whitlock another electrostatic generator.

3.5. *The Agrimeter*

If, instead of the plate being fixed and the cover moving, the plate moves from an open to a covered position, so that it is earthed in the open position and connected to a collector in the covered position, then the collector receives the negative charge. If this is done continuously, the collector receives an intermittent charge. This charge can be used either to charge some form of electrometer or to provide a current.

The method was first used by Russeltvedt (1925) and later by Workman and Holzer (1939) with electrometers and by Chalmers (1953c) with a galvanometer.

3.6. *Change of Field with Height*

One of the most important results of the measurement of fields is that the field in fine weather is nearly constant through the first few metres of height above the ground, so that there is little space charge in this region; measurements by Watson (1929) and Scrase (1935) have shown that the field at the earth's surface differs only by a few per cent from the average over the first metre of height. This result is confirmed by direct measurement of the space charge (e.g. Brown 1930).

Unless there is a current which is not a conduction current (for example a convection current), the conclusion to be drawn from the constancy of F must be that λ is constant in the first metre of height.

But, while at a height of about one metre the conduction current is carried by ions of both signs, with, as measurement shows, approximately equal shares, at ground level it is clear that it is only positive ions which can carry the current.

This led to the idea that one might expect the conductivity at the earth's surface to be only about half that at one metre.

3.7. *The Idea of Convection Currents*

If the fields at ground level and at one metre are nearly equal and yet the conductivities are in the ratio about 1:2, it follows that the conduction currents must be in the ratio of 1:2 also, and so there must be some other current, presumably a convection current, carrying positive current upwards at one metre to about the same extent as the conduction current carries negative charge upwards. This idea gained considerable support at one time, although it was never found possible to explain what this convection current could consist of, in the absence of appreciable space charge.

3.8. *Direct Measurement of Conduction Current*

Clearly, the most satisfactory way to measure the conduction current at the earth's surface is to isolate a portion of the surface and measure the charge arriving there in a given time. Unfortunately this gives, not only the charge arriving from the air, a true current, but also the charge corresponding to the change of field, and hence of bound charge, a 'displacement current'. To measure the conduction current alone, some method must be found of eliminating the change of bound charge; Simpson (1910) and Chalmers and Little (1947) took averages over long periods of time and assumed that field changes would balance out; Kasemir (1952) arranged to have a very long time constant (about 1 hour) so that field changes would be smoothed out; and Scrase (1933) attempted to compensate for field changes by applying the field to the measuring apparatus through a suitable condenser and so obtaining an effect equal and opposite to the change of bound charge. In its simple form this would have achieved the desired result, but it did not work satisfactorily, and Scrase had to put a wire net, connected to his potential equalizer, over the plate, and so in effect he measured the conductivity for positive ions close to the surface, under conditions as close as possible to the natural, and from conductivity and field obtained the air–earth current.

If the collecting plate is covered before the measurements begin, uncovered for measurement and then covered again, the total charge received is that of the conduction current alone, provided that there is no leakage during the time. But, unless suitable adjustments are made, the plate is not at earth potential during the collection and conditions are therefore wrong. Such a method has been used by Wilson (1916) with a capillary electrometer and by Nolan and Nolan (1937) in their comparison of the currents as measured by the direct and indirect methods; but these were for visual observations, and it would be difficult to adapt the method for continuous measurements.

3.9. *Indirect Measurement of Conduction Current*

The indirect method of measuring the conduction current consists of measuring the field F and the conductivity λ at the height concerned.

Results for the value of the conduction current vary from place to place owing to local conditions, and it seemed from earlier results that the idea expressed above was borne out that the conduction current at 1 metre, measured necessarily by the indirect method, was about twice that at the ground, measured by the direct method.

But the first case of measurements by both methods at the same place, that of the Nolans at Glencree (1937), showed that the two values of the conduction current, instead of giving a factor of 2:1, differed only by 12%. This meant, first, that the convection current is only a fraction of the conduction current, and, second, that the positive conductivity at the ground must be considerably greater than that at one metre.

It would be very desirable to verify the result of Nolan and Nolan at other places and over longer periods of time, preferably with continuous recording.

3.10. *Variation of Conductivity with Height*

The conclusion that the results of the Nolans must involve a greater positive conductivity at the ground than higher up was verified by Hogg (1939), who found that the total atmospheric conductivity remains nearly constant from the ground up to 1 metre, but that the proportion due to the positive ions increases from about 50% at 1 metre to 100% at the ground.

While this result of Hogg's carries the argument a stage further, it is by no means obvious how such a change of conductivity can come about. Chalmers (1946) showed that it could arise if the rate of production of ionization is markedly higher close to the ground than higher up, and showed that, in a case chosen for illustration, no significant space charge would arise.

Hess and O'Donnell (1951) have recently given experimental verification of the idea that the rate of ionization varies with height, and the matter appears to be satisfactorily rounded off. But it remains a matter of comment that it seems to be only in the nature of a lucky accident that the rate of variation of ionization with height and the other conditions conspire to give very small space charge.

It should be pointed out that the apparent contradiction between the results of Hogg (1939) and of O'Donnell (1952) for the variation of conductivity with height is accounted for by the difference of conditions, since Hogg used an open site while O'Donnell's measurements were made under trees; the author (Chalmers 1953 a) has discussed this in more detail.

3.11. *Measurements of Field on Mountains*

The two factors which influence the variation of the fine weather field at any point with time of day are (*a*) the potential of the ionosphere and (*b*) the local conductivity. The local conductivity is dependent on the time of day mainly by reason of different extent of interchange of air between different layers of convection, the process often described by the German word 'Austausch'.

If 'Austausch' does not extend to a station on a mountain, then the field will show a maximum at about 19 h G.M.T., corresponding to the maximum of the potential of the ionosphere. But, on the other hand, if 'Austausch' does extend that far, then the variation with time of day will be different.

Israël, Kasemir and Wienert (1951) have made measurements at the Jungfraujoch and find results which show that 'Austausch' reaches to this height (3472 m) in the summer (June and July) but not in the autumn (October); in the summer the relations and variations of field, current and conductivity at the Jungfraujoch are similar to those at any land station, but in the autumn the similarity is to ocean or polar measurements.

This type of measurement may well be the forerunner of the application of methods of atmospheric electricity to the study of more purely meteorological problems.

§4. Point Discharge Currents

4.1. *The Process of Point Discharge*

If the field strength is sufficiently great there can be point discharge, or 'corona discharge', in the neighbourhood of a raised point connected to the earth. In a limited volume round the point, ionization by collision can occur and the ions of one sign carry charge into the point, while those of the opposite sign move away in the field.

Measurements with trees (Schonland 1928) and with prepared points of similar height (Wormell 1927, etc.) have shown that point discharge occurs naturally when fields approach 1000 v m^{-1}.

Investigations of corona discharge in the laboratory have been made by Trichel (1938) and others and have shown that the current consists of a series of pulses of similar size, the frequency depending on the field. Atmospheric point discharge has also been found to consist of pulses (Smith 1951).

4.2. *Distinction between Atmospheric and Laboratory Point Discharge*

In laboratory experiments on point discharge the volume in which ionization by collision occurs is usually of dimensions comparable with the distance between the electrodes, but for atmospheric point discharge the 'electrodes' are the point, at earth potential, and the cloud. Thus in laboratory experiments the ions that have moved away from the point quickly move to the other electrode and have no effect, whereas in the atmosphere they can affect matters very much by acting as a space charge and so reducing the field from that which would occur in their absence.

Thus results of laboratory experiments cannot in general be carried over to atmospheric phenomena, and many of the investigations of atmospheric point discharge must be carried out in the atmosphere, where the conditions are very much less under control.

4.3. *Conditions of Point Discharge*

There are various conditions of point discharge which are of interest in regard to atmospheric point discharge. First, there is discharge from a tree or a pointed conductor of about the same height as the trees in the neighbourhood, so that the point may be considered as one of a number of similarly situated points. Second, there is the single point or tree which is considerably higher than its neighbours. This may give natural point discharge for fields somewhat lower than in the previous case. Next there is a point connected to earth and at a sufficient height to give point discharge in fine weather when the field reaches only 100 to 200 v m^{-1}. Such a point must be at least 30 m high and well away from trees, etc.

The next case is that of the alti-electrograph where a balloon carries points below itself at two different levels and the system reaches a potential relative to its surroundings such that the same point discharge current flows through the upper and lower points. The actual alti-electrograph is complicated by the fact that the balloon is continually rising.

Finally one might imagine a balloon with a sufficiently great distance between the points to give discharge in fine weather. In addition to all these, there is the question of multiple points instead of single points in every case. It is clearly

not justifiable to use the results of point discharge in one set of conditions and apply them directly to another set.

4.4. *Relation between Point Discharge Current and Field*

The relation between the current I through a single point, of roughly the same height as the trees nearby, and the field F at the earth's surface has been measured by Whipple and Scrase (1936), Chiplonkar (1940) and Hutchinson (1951). These show general agreement with a formula of the type

$$I = a(F^2 - M^2)$$

where a and M are constants which must depend on the height of the point and the arrangement of other points nearby.

For a single point attached to a captive balloon and considerably higher than other points nearer the ground, e.g. trees, Chalmers and Mapleson (1953) have measured the point discharge current in fine weather, the field nearby and the wind speed, as well as the height of the point. Their results show that the current depends on the product of the field F at the earth's surface and the height H which gives the potential V that would exist at the height of the point in the absence of the point, provided that there is no space charge in the region below the point. Assuming that the results can be put in the form $I = AV^{\alpha}W^{\beta}$ they found $\alpha = 1{\cdot}786$, $\beta = 0{\cdot}244$. The sum $\alpha + \beta$ then is $2{\cdot}030 \pm 0{\cdot}052$, while theory (see below) gives $\alpha + \beta = 2$. Assuming this, the result is finally $I = 0{\cdot}021V^{1{\cdot}77}W^{0{\cdot}23}$ where I is in amps, V in kv and W in km hr^{-1}. It was not found possible to introduce a constant corresponding to M, although there is no current when V is too small.

Lutz (1941) and Hutchinson (1951) have found that the point discharge current and field are not exactly proportional during periods of rapid change. Lueder (1943) and Hutchinson (1951) have explained these results as due to space charges in the region between the height of the point and the earth, produced by point discharge at lower points. Hutchinson (1951) has also used a similar basis to explain deviations from the simple Whipple and Scrase formula.

4.5. *Theoretical Considerations*

Chalmers (1952b) has given a theoretical derivation of the relation between point discharge current and field for a point which is one of a series of points in rectangular array, and obtained a result agreeing with the relation of Whipple and Scrase (1936). This derivation is based on the idea that the space charge of the ions moving away from the point just limits the volume in which ionization occurs to produce enough current to maintain the conditions. It is assumed that the space charge ions remain above the point and are not affected by wind; but the effect of wind would not be important since space charge ions removed from above one point would be nearly replaced by those from above others of the points.

Chalmers' derivation applies only to one of a series of points in rectangular array, but has been adapted to the case of a single point by Chalmers and Mapleson (1953) taking into account the removal of ions by the wind. They obtained $I = AV^{\alpha}W^{2-\alpha}$ where α is undetermined by the theory. This agrees well with experiment.

4.6. *Space Charge from Point Discharge*

When point discharge occurs, ions of one sign send charge into the point, while ions of opposite sign move upwards, forming a space charge. As was first pointed out by Wilson (1925), this space charge must involve an increase of field on rising and, as Whipple and Scrase (1936) calculated on reasonable assumptions, the field below the cloud could be very many times the field at the earth's surface. The measurements that have been made of the fields below clouds, e.g. the alti-electrograph measurements, have not shown this increase of field, and various attempts to overcome this difficulty (Chalmers 1939, 1944) have not been successful. Malan (1952) has suggested that the effect of wind in removing the space charge might account for the discrepancy.

4.7. *The Alti-Electrograph*

Simpson and Scrase (1937) used pole-finding paper to measure the currents flowing through points at two levels below a freely ascending balloon. From the measurement of the traces they found the currents and endeavoured to estimate the fields, assuming $I = aF^3$ (M can be neglected) and using values of a found by measuring F as the balloon started to ascend. A correction was made for change of a with pressure. The fields so obtained were surprisingly small and did not show the expected increase with height.

The conditions of these measurements are so different from those for which the formula has been established that it clearly needs further investigation before it is possible to confirm or deny that the formula holds and, if it does hold, that a is independent of the height of the balloon.

It is difficult to establish any theoretical formula since the balloon is rising and it is, at the same time, producing ions which move upwards in the field. There are currents through the points at the two levels, and the potentials of the wires must adjust themselves so that both currents are equal. Below the balloon and wires conditions are very complex; in a negative field positive ions will be coming up from the points at the ground, negative ions will be moving down from the lower point, and there may also be positive ions from the upper points moving downwards relative to the balloon, if the rate of ascent is sufficiently great.

Thus it would appear that the use of the Whipple and Scrase formula is scarcely to be relied on; it may even not give relative values of field at different heights since it may be that the current–field relation depends on the height of the balloon.

Though there is as yet no theoretical or experimental basis, it would seem that the alti-electrograph results could best be brought into line with what would be expected by the idea, suggested in the author's book, that the current through the alti-electrograph represents, at all heights, the existing point discharge current density over some area.

4.8. *Multiple Points*

Chiplonkar (1940) found that, in the same field, four neighbouring points gave together somewhat less point discharge current than a single point. On the other hand, Belin (1948), in laboratory experiments, used 7 and 13 points and found the same amount of current per point in the same field. Chalmers and Mapleson (1953), in measurements of point current from a balloon, found that eight points gave together less current, in the same field, than a single point.

It seems clear that the discrepancy must be due to the difference in conditions between atmospheric and laboratory investigations. Whereas the separations of

the points in the atmospheric investigations were certainly small compared with the scale of the whole, in Belin's case the separations may well have been of the same order as the distances concerned (Belin gives no actual dimensions).

4.9. *Total Point Discharge Current from Cloud*

It might well be expected that the total point discharge current from the whole area beneath a cloud would depend upon the nature of the surface beneath the cloud and, in particular, on the nature and number of points. But Chalmers (1952 c) has shown that this is not so, and that the total point discharge current beneath a cloud depends very largely on the rate of separation of charge within the cloud and very little on the surface beneath. If it is assumed that there is a definite rate of separation of charge in a cloud, giving a definite vertical electric current of charge separation, then the fraction of this current which is dissipated by currents outside the cloud, i.e. to the ionosphere and to the earth, is very little dependent on conditions near the earth, such as the nature and number of discharging points. Hence, wherever the cloud may be, the current through it remains nearly the same since, as Schonland (1928) has shown, point discharge is the most important of the currents between earth and cloud and, therefore, lightning and precipitation currents need not be considered.

4.10. *Effective Separation of Discharging Points*

In discussing some features of point discharge, e.g. the current between cloud and earth, it is of more importance to obtain the total point discharge current over an area than the current through a single point. A single point or tree can often be considered to be one of a rectangular array of points spread regularly, or, in other words, the current through the point is the total current over a certain area. The problem then is to determine this area, or the effective separation of points similar to the one used.

Schonland (1928) used an actual tree, typical of the locality, and mounted on insulators. If his tree gives the same discharge current as in the natural state, then the effective separation of points is easily obtained. But, as Dr. Mapleson pointed out to the author, Schonland's photograph suggests that the point is better exposed than in natural conditions. Other workers have used special points and have only been able to obtain the effective separation by estimation or by indirect methods.

On the assumption that rain obtains its charge from the point discharge space charge, Simpson (1949) found an effective separation of 22 m for Kew to compare with Whipple and Scrase's estimate of 25 m. Chalmers (1951), by the same method in more detail, obtained a value of 14·5 m, and from the results of Hutchinson and Chalmers (1951) a separation at Durham of 6·1 m.

Smith (1951) considered the variation of field after a close lightning discharge to be due to the removal of space charge, and was thus able to deduce the point discharge current density.

Chalmers (1953 b) has obtained a separation of 11 m for Kew by assuming that the total point discharge current below a thundercloud must equal the value of 0·5 amp found for the average current through a cloud (Gish and Wait 1950).

If there are a number of discharging points at different heights, then it is not possible to define a definite effective separation, and this will vary with the strength of the field.

The indirect methods used to obtain the effective separation of points have led, for Kew, to values less than the early estimate, and so have increased the difficulties mentioned above (§ 4.6) in regard to the discrepancy between the fields measured below clouds and the fields expected from space charge considerations.

§ 5. Precipitation Currents

5.1. *Importance of Precipitation Currents*

The measurement of precipitation currents has two reasons for importance in atmospheric electricity. In the first place it may be that precipitation plays an important part in the maintenance of the earth's negative charge, and therefore it is important to measure the charges brought down by precipitation. Secondly, it is generally recognized that precipitation particles must play an important part in the processes of charge separation in clouds, and when they reach the ground we have these particles accessible to measurement; so we may hope that measurements of precipitation currents can help to elucidate the mechanism of charge separation in clouds.

It may be stated at once that, in fact, results of measurement of precipitation currents do not notably contribute either to the problem of the maintenance of the earth's charge nor to the problem of the separation of charge in clouds, but it is only by carrying out the measurements that it has been possible to make such an assertion and to falsify the hope that this might be a way of progress in the main problems of atmospheric electricity; and the results have posed other problems as to the origin of the charge on the precipitation.

5.2. *General Results*

The general results up to date have been summarized by Chalmers (1949), and the more recent work has not altered these conclusions. The most important results are the following: precipitation currents range from 10^{-16} A cm^{-2} up to 10^{-12} A cm^{-2}, with exceptional cases of greater values. The currents more usually bring a positive than a negative charge to the earth, both for thunderstorm rain and for continuous rain, but there is an excess of negative charge for showery non-stormy rain. The general excess of positive charge shows that precipitation cannot be the way in which the negative charge on the earth is maintained, but rather that precipitation currents add to the fine weather current in bringing positive charge to the earth.

But it must be recognized that the results quoted are perhaps not typical of the whole earth, since they have been obtained only at a comparatively few places and those mainly in temperate latitudes. Measurements in arctic or tropical precipitation might completely alter the conclusion above.

5.3. *Relation with Field*

The earliest workers, e.g. Elster and Geitel (1888), had found an inverse relation between rain current and field, but later workers had not found it except in exceptional cases, e.g. Chalmers and Little (1947), until Simpson (1949) produced some conclusive evidence in favour of such a relation.

Simpson divided his results into two categories, cases of fields between ± 1000 V m^{-1}, in which there was no point discharge, and fields outside the limits of ± 2000 V m^{-1}, when point discharge operates.

For the region of lower fields Simpson found the law

$$q = -0{\cdot}0145\,(P-4)$$

where q is the charge, in e.s.u., per unit volume and P is the surface field in v cm^{-1}. Simpson suggested that the '4' represents the normal field in fine weather and so $P-4$ is the displacement from the normal field. Smith (1951) has suggested that Simpson's result could occur if the rain fell with little or no charge, and acquired a charge by splashing, the corresponding charge of opposite sign (negative) going into the air and giving rise to a decrease of the field from its normal value. An objection to Smith's idea comes from the fact that a predominantly positive rain current is found even when the receiver is shaped so that splashing effects could not occur, and that single drops show positive charges when no splashing can occur.

When point discharge occurs, Simpson found that the rain current is proportional to the point discharge current but of opposite sign, the factor of proportionality depending upon the rate of rainfall and tending to a limit for the highest rates.

Since it seems that the sign and the magnitude of the charges on rain depend on the field, it follows that the rain charges can give little, if any, fresh information about the process of separation of charge in clouds. However, the origin of rain charges poses a fresh problem which has by no means yet been completely solved.

5.4. *Single Drops*

In investigations aimed at finding out the origin of the charges on rain, measurements of single drops may give more information than those of a volume of rainfall.

Earlier work on single drops by Gschwend (1922) and by Chalmers and Pasquill (1938) has been extended by measurements by Gunn (1949), Federov (1951), Hutchinson and Chalmers (1951), Smith (1951) and Banerji and Lele (1952). In addition to drop charges, all except Federov have measured drop sizes, and the last three have also measured the field. Federov found that some drops carry charges considerably smaller than had previously been measured, values going down to 10^{-5} e.s.u., while Banerji and Lele (1952) obtained charges up to over 10^{-1} e.s.u. Gunn's measurements were confined to thunderstorms.

Hutchinson and Chalmers found definite evidence of the inverse relation of drop charge and field, with 629 cases of opposite sign as against 286 of the same sign. For 173 drops during periods when point discharge was measured there was found a definite correlation between drop charge Q and point discharge current I, the two being of opposite sign, and there is probably a dependence of $-Q/I$ on a, the radius of the drop. Rate of rainfall was not measured, but if it is assumed that the drops measured were of the average size for the rate of rainfall, and making use of Best's (1943) relation between average size and rate of rainfall, it is found that the results agree well with those of Simpson (1949).

Smith (1951) measured the charges of a large number of drops in each of several periods of rain. He found that, while the smaller drops carried charges of sign opposite to the field, larger drops often carried charges of the same sign as the field. On adding the charges in one period of rainfall, Smith found the total to be often of the same sign as the field, but pointed out that small drops, with opposite charge, would frequently be missed and, if included, might give again the inverse relation.

These results of Hutchinson and Chalmers and of Smith show a definite discrepancy which could not easily be resolved except by assuming that Hutchinson and Chalmers seldom measured the drops larger than the average, which Smith found to be of opposite sign, but measured mainly average drops.

Smith's results may explain why it is that the inverse relation of drop charge and field has so often been missed. In a great many of the measurements of rain charges great care has been taken that the receiver should be shielded electrostatically so as to avoid effects of field changes. This shielding may well have had the effect of preventing the smallest drops from reaching the receiver when there was much wind. So these small drops would not make their contribution to the total charge and the inverse relation would be missed; Chalmers and Little (1947) found the relation on certain occasions, with a fully exposed receiver, and Simpson (1949), with a receiver not as well shielded as those of earlier workers, found the effect very markedly.

5.5. *Interpretation of Charge–Field Relation*

There are two main ways in which the relation between rain charge and field could be interpreted. In the first place it could be thought that there is some process which gives a charge, usually positive, to the rain, the opposite charge remaining in the cloud, or perhaps in the air below; this latter charge would then give rise to a field at the ground of sign opposite to that of the rain charge. This is the interpretation preferred by Scrase (1938) and assumed by Banerji and Lele (1952).

On the other hand it may be that the field is first set up and the raindrops acquire their charges by the capture of ions moving in the field, according to the theory first put forward by Wilson (1929). This is the theory used by Simpson (1949), by Chalmers (1951) and by Smith (1951).

As was pointed out by Chalmers and Little (1940), there is a clear distinction between the two methods of interpretation as regards the total current carried between ionosphere and earth. If we use the first interpretation, then there is a process of separation of charge giving, usually, positive charge to the rain and leaving negative charge aloft. Unless there is also some other process of separation of charge, this must mean a resultant positive charge coming downwards. If conditions are anything like steady then there must exist this same resultant current, positive downwards, at all levels. On the other hand, if the rain acquires its charge from an ionic current it follows that the rain current can never be as large as the ionic current and so, usually, the resultant current brings negative charge downwards.

Simpson's (1949) results, at times when the field was strong enough to give point discharge, showed the rain current to be proportional to the point discharge current, the factor of proportionality depending on the rate of rainfall and tending to a limit for high rates. This is easily understood if the rain gets its charge from the ions in the point discharge current, the limit corresponding to the case when all the point discharge ions are captured.

That this is the correct interpretation for the case of cumulo-nimbus clouds is rendered much more probable by considering the direction of the resultant current within the cloud, where it certainly brings negative charge down, and above the cloud, where Gish and Wait (1950) have also shown that there is a negative current downwards.

5.6. *Origin of Rain Charges*

Chalmers (1951) has been able to give an approximate quantitative account of the average values of the rain charges as obtained by Simpson and by Hutchinson and Chalmers, under conditions when there is point discharge, by the following mechanism. Some process of charge separation in the cloud gives the precipitation particles a charge of one sign, let us say negative, while charges of opposite sign rise in the cloud. There is thus a negative charge in the base of the cloud, and this produces a negative field at the ground; if this is sufficiently strong there will be point discharge at points near the earth, and positive ions will travel upwards. The falling raindrops will capture these ions and, in general, will become positively charged before they reach the ground, thus giving the inverse relation. Numerical calculations, including the space charges of the ions and the drops, show that the general results of the observations agree with the picture given.

Smith (1951), working on somewhat similar lines, has shown that it is to be expected that the larger drops will retain their original sign of charge, while small drops under the same conditions change sign, thus giving an explanation of Smith's own results.

5.7. *Rain Charges in Low Fields*

The interpretation of the results for the charge on rain in fields too low to give point discharge appears to be different from that for higher fields.

Simpson's (1949) results show that when there is no field the rain carries a positive charge and when the rain carries no charge the field is positive. Thus, if the only vertical currents present are the rain current and the conduction current, in these two cases one of the two is absent, and the other brings down a positive charge.

It thus appears as if the case of continuous rain with low fields can be interpreted as the rain obtaining its charge high up and leaving behind the charge of opposite sign.

5.8. *The 'Mirror Image Effect'*

Simpson (1949) not only found that the rain current and field were opposite in sign usually during periods of steady conditions, but also that they frequently changed sign simultaneously in opposite directions, so that the course of the rain current appeared as the mirror image of the course of the field.

A raindrop takes several minutes to fall from the cloud to the earth, so that one might expect that, if the raindrop obtains its charge in or near the cloud, and this depends on the field existing at the time, then there might well be expected to be a delay between the field change and the arrival of a raindrop with a different charge, a delay which is not found. But it must be realised that a change in electrical conditions at cloud level may not show itself immediately at the ground, as there will have to be adjustments of the space charges in the atmosphere; such adjustments will take place with a 'relaxation time' t of $1/4\pi\lambda$ where λ is the conductivity of the atmosphere (in e.s.u.) and t has a value of 5 to 30 minutes, depending on the locality, if the conductivity is that of fine weather.

If, however, the raindrop obtains its charge near the ground by a process depending on the field, then the mirror image effect would be easily explained. This would seem to agree with the idea of the raindrops acquiring their charges by ion capture; but the calculations of Chalmers (1951) show that, if ion capture is to explain rain charges, it is ion capture at a considerable height that is most important, and the simple solution to the problem of the mirror image effect cannot hold.

If it could be considered that the change of field observed is not due to changes taking place in the cloud charges, but rather to the motion, over the observer, of a cloud carrying different charges in different parts, then it can be suggested that the cloud, the rain and the ions all move at roughly equal horizontal speeds, and the mirror image effect would hold at whatever height the fundamental charging process for the rain might occur.

Another way of discussing the mirror image effect is the following: suppose that some process charges the raindrops and that any line of force with one end on a drop has its other end in the cloud. Then, as soon as a drop reaches the ground, the line of force must also end on the ground, giving a field of opposite sign to the rain charge.

5.9. *Fields and Currents in Fog and Mist*

In fog or mist the usual effect is just what would be expected from a decrease of conductivity, that is to say there is a slight decrease in the conduction current density and a considerable increase in the field. Israël and Kasemir (1952) have found that, on the average, the conductivity is reduced to about one third of its usual value.

Chalmers (1952a) has found that, close to high-tension overhead cables, there is a negative field to leeward in mist or fog, with the usual positive field to windward. This must be ascribed to negative ions released at the insulators by the brush discharge. Under suitable conditions the effect may be evident for several miles from the cables.

Scrase (1933) and Israël and Kasemir (1952) have found cases where the field and the current are of opposite sign, so that there must be some other current in addition to, and of opposite sign to, the conduction current; in such cases, though there is no precipitation detectable by a rainfall recorder, very often the fog is wet and there is probably very weak precipitation carrying down usually a negative charge.

In measurements in fog and mist on mountain tops, more accurately inside clouds, Israël and Kasemir (1952) found rather different results from those on plains; here there is a considerable decrease in the vertical current with little change in field, giving again a decrease of conductivity to about one-third of its usual value. This is explained as differing from the results for plains in that the current can go round the poorly conducting cloud, while this is not possible for a widespread ground fog.

§6. Maintenance of the Earth's Charge

6.1. *Fine-Weather Current*

In fine weather the field at the earth's surface is positive, so that there is a negative charge on the surface. Owing to the conductivity of the atmosphere there is a current of positive charge to the earth in fine weather, and this would neutralize the charge on the earth in a time which is from 5 to 30 minutes, depending on the actual value of the conductivity. But, in spite of this, the negative charge on the earth remains constant from year to year, nearly constant from day to day, and does not alter very much even from hour to hour.

If we consider only those places where local pollution does not affect the field, and consider measurements at sea to be representative of the fine-weather parts of the earth, then we can take Mauchly's (1923) value of 126 v m^{-1} as the average fine-weather field. Mauchly's value for the conductivity was

$2{\cdot}3 \times 10^{-4}$ e.s.u. $= 2{\cdot}5 \times 10^{-16}$ ohm^{-1} cm, which gives a time of 5·8 minutes for the current arriving at a portion of the earth's surface to balance the charge already there. This gives an air–earth conduction current of $3{\cdot}2 \times 10^{-16}$ A cm^{-2}. Neglecting areas of bad weather, the fine-weather current to the whole earth is thus $+1630$ A on the average, which must be balanced by an equal negative current. Gish and Wait (1950), using slightly different data, obtained a value of 1800 A.

6.2. *Variation with Time of Day*

Many measurements of field have shown that, where local pollution has no effect, the field shows a variation with the time of day which is the same all over the world and depends, not on local time but on world-wide time giving a maximum around 19 h G.M.T. At the same time some of the measurements have shown that there is no significant variation of conductivity with time of day, so that it can be concluded that the air–earth current shows the same variation as the field.

Then it must follow that the difference lies in a difference in the potential of the ionosphere from one time of day to another. Calculations of the resistance of a column of air up to the ionosphere, based on conductivity measurements at different heights, give a value of around 10^{21} ohms and a potential of the ionosphere of 3 to 4×10^5 v on the average.

The deviations from the mean value of field, for different times of day, amount to around $\pm 20\%$, so that the potential of the ionosphere varies to this extent.

6.3. *Correlation with Thunderstorms*

It was first suggested by Appleton (1925) that the maximum of field at 19 h G.M.T. might be correlated with a maximum of thunderstorms over the whole world at that time. This idea was explored by Whipple (1929) and by Whipple and Scrase (1936), using available statistics of thunderstorm frequency, and they found good correlation, not only of the time of maximum but also of the general shape of the curve and of the ratio of maximum to minimum.

Wichmann (1951), however, pointed out that Whipple included sea thunderstorms equally with land storms and that the current to the earth from sea storms might be expected to be much smaller than from land storms, since there would be much less point discharge at sea than on land, and point discharge, as shown by Schonland (1928), provides the greatest part of the current to earth during a storm. If the sea storms are left out, Whipple's correlation is much less good, particularly in regard to the ratio of maximum to minimum. But Chalmers (1952 c) has shown that, in spite of the absence of points, the total current to earth beneath a cloud is very nearly the same whatever may be the nature of the earth's surface below (see § 4.9), and thus Wichmann's (1951) criticism no longer holds.

6.4. *Effect of Induced Charges*

Chalmers (1953 e) has pointed out that the typical storm cloud induces a positive charge on the earth and a negative charge on the ionosphere and that, at the time of thunderstorm maximum, these induced charges will be largest, so that there will be made available larger negative charges on the earth and positive on the ionosphere for the fine-weather parts of the earth at this time. Thus even without the transfer of negative charge to the earth during storms, there would still be a maximum of fine-weather field at the time of thunderstorm maximum due to induced charges.

From the observational result of Gish and Wait (1950) that the conductivity of the atmosphere above the clouds is not different from that at the same level in fine weather, Chalmers (1953 d) has shown that the total charge on the inner side of the ionosphere is always zero. If we take a sphere which is large enough to include all electrically active clouds as well as the earth, and consider the volume between this and the ionosphere, the conductivity depends only on the height and not on the presence or absence of clouds below. If there should be more electrostatic lines of force crossing the volume in one direction than the other at any time there will be a greater current in that direction, and this would inevitably reduce the difference in the number of lines of force, so that there would soon be no such difference and hence equal quantities of charge of either sign on the inner side of the ionosphere.

Then the actual amount of charge on the ionosphere giving lines of force proceeding to the earth must be equal and opposite to that giving lines proceeding to clouds, and the current to the earth must equal the current from the clouds, giving a very simple account of the diurnal variation of current and hence of field.

6.5. *Currents below Storms*

The idea, originally propounded by Wilson (1920), that it is storms which are responsible for the maintenance of the earth's fine weather surface charge, is supported by the correlation of field maximum with thunderstorm maximum, but it must stand or fall by measurements of the actual charges brought to the earth during storms.

Wormell (1930) first attempted to draw up an 'electrical balance sheet' for an area of the earth's surface, and found, from his estimates for Cambridge, that there might be an excess of negative charge reaching the earth. Chalmers (1949, p. 121) has used results for Kew and reaches a similar conclusion. But over the oceans, where the fine-weather current is larger than at Cambridge or Kew, it may well be that the balance is closer. In tropical regions, where storms are more frequent, it may well be that there is a great excess of negative charge, while in deserts and in polar regions, where there are few, if any, thunderstorms, there will be an excess of positive charge reaching the earth.

Another way of regarding the problem is to consider what charge is required from each storm if Wilson's theory is to be borne out. Brooks' (1925) statistics give about 1800 storms in existence at any one time, so that if there is to be 1630 A for the fine weather current there must be about 0·9 A per storm. This neglects any transfer of charge during non-stormy rain, and Wormell (1927, 1930) has shown that point discharge currents are nearly as strong in some showers without thunder as in genuine thunderstorms. Also, as Gish and Wait (1950) have pointed out, Brooks' estimate of 1800 storms at one time may well be somewhat low.

6.6. *Currents above Thunderclouds*

The final confirmation of the idea that it is storm clouds which provide the currents maintaining the earth's charge has come from the measurements of Gish and Wait (1950), who flew an aeroplane over the tops of thunderclouds, measuring the field and the conductivity; by integrating over the whole area above the storm they were able to estimate the total vertical current above each storm. Their results showed that the current always takes positive charge upwards, with currents usually from 0 to 1·4 A, but in one exceptional case, in which the aeroplane flew

through the upper part of the cloud, a value of 6·5 A was obtained. The average value was 0·8 A, or, if the exceptional case is excluded, 0·5 A.

Gish and Wait took the fine weather current as $3{\cdot}5 \times 10^{-16}$ A cm^{-2}, giving a total current of 1800 A, so that there would have to be 2200 to 3600 storms at once. With the qualifications to Brooks' result of 1800 storms, as mentioned above, it does not seem unreasonable to accept values of 2200 or 3600.

There seems now no reason to doubt that it is thunderstorms which maintain the earth's fine-weather field against the conduction current, and it may safely be said that this problem can be regarded as solved.

§ 7. Thunderclouds

7.1. *Development of Clouds*

Modern investigation of thunderstorms has shown that the essential unit in a thunderstorm is the 'cell', and that a storm consists of one or more cells in various stages of development (Byers and Braham 1949).

The cell shows three distinct stages in its development. In the first or 'cumulus' stage the vertical air currents are everywhere upwards and may reach values of 15 m sec^{-1} and over. No precipitation is falling out of the cloud during this stage.

The second or 'mature' stage commences when the precipitation particles are large enough to fall out of the cloud. As the precipitation increases so it carries air down with it, though strong up-currents also still exist amounting to as much as 30 m sec^{-1}. It is in this stage that lightning occurs and that the other phenomena of the thunderstorm reach their greatest intensity.

In the final or 'dissipating' stage the down-current spreads over the whole cell and all the effects decrease in intensity. The complete lifetime of a cell is usually somewhat under half an hour. The diameter of a cell is of the order of a few kilometres.

7.2. *Methods of Electrical Investigation*

The large and rapidly varying fields below thunderclouds show decisively that there must be large charges resident in the clouds. But it is a matter of elementary electrostatics that even a very large number of simultaneous measurements of field at different places on the surface of the earth cannot give any definite solution to the problem of the magnitudes and positions of the charges in the clouds. A very simple example will suffice to show the truth of this statement: suppose that there is a uniform field of a certain magnitude over a wide area; then all we know is that somewhere higher up there must be a layer of charge equal and opposite to that on the earth's surface, but there is no means from surface measurements of telling how high this is, nor whether there are other layers of charge higher up.

Some further progress can be made by measurements of the field changes due to lightning flashes, since in this case we know we are concerned with only a single charge moving from one place to another, rather than the effects of an indefinite number of charges. But, in order to obtain results that are certain, it is necessary to know whether the flash is to earth or not and how far away it is, and also to assume that it is nearly vertical. The theory underlying these results has often been given and will not be repeated here.

In order to obtain more definite results it is necessary to make measurements at points above the earth, and this involves the use of balloons or aircraft.

Balloons can either carry apparatus for recording, to be recovered later, or else can transmit readings by radiosonde; a disadvantage of balloons is that they are subject to winds and so may not arrive exactly where it is desired. Aircraft, on the other hand, have the disadvantage that they become charged themselves, disturbing the readings, and there is considerable danger to an aircraft in a violent storm.

7.3. *Results of Field and Field Change Measurements*

The general conclusion can be reached, from measurements of the field beneath a thundercloud and of field changes brought about by lightning, that the typical thundercloud has an upper positive charge and a lower negative charge, and lightning flashes bring down this lower charge. This general picture is only statistically true and seldom holds for any single storm in detail, so that there are quite frequently considerable divergences from this simple structure.

At one time Simpson (1926) believed that lightning brought positive charge downwards because of the direction in which a lightning flash branches, but this was explained by the experiments of Schonland and Allibone (1931), who showed that the direction of branching depends, not on the signs of the charges, but on the concentration of the lines of force.

Wormell (1953) has discussed, in some detail, results of field changes and the correlation of photographic and electrical measurements. He has made some interesting deductions as to the magnitudes of the charges undergoing separation, about 1400 coulombs, and the weight of precipitation concerned, over $1{\cdot}4 \times 10^{10}$ g.

7.4. *Currents in Thunderclouds*

Measurements of the field changes produced by lightning flashes gave values of about 30 coulombs for the amount of electricity carried by the average lightning flash (Bruce and Golde 1942). After a flash the electric field recovers and often tends to return to its original value; from the initial rate of recovery there appears to be a current in the cloud of several ampères, but this decreases as the potential difference within the cloud builds up, and there is set up an opposing current within the cloud.

If we consider a cloud in which the fields are nowhere quite sufficient to produce lightning but in which the conditions are steady, with the processes of dissipation of charge within and outside the cloud just balancing the process of separation of charge within the cloud, then we shall have a cloud whose electrical activity is somewhat less than that of a cloud which produces lightning discharges, but which can be discussed more easily since conditions are steady. In such a cloud the rate of separation of charge must be some ampères, with dissipation within the cloud balancing a considerable part of this, the remainder being balanced by currents above and below the cloud, of values about $\frac{1}{2}$ amp.

7.5. *Currents above and below Clouds*

If conditions are steady, then the total current through the cloud must be the same at all levels, though the manner in which it is carried may differ from one level to another. The results of Gish and Wait (1950) have shown that above the cloud there is a current of less than one ampère carrying positive charge upwards; this is a simple conduction current, carried by ions of both signs.

Below the cloud there must be a current of the same size again carrying positive charge upwards; but this is not so simple as the current above the cloud,

since it must comprise both an ionic current and also a current of precipitation particles; if we are considering steady conditions with no discharges we can neglect lightning currents. The precipitation current can be measured directly, and this has often been done; but the ionic current, in the high fields present, is a point discharge current, and to measure this as a current density or as a total current beneath the cloud requires a knowledge of the effective separation of the discharging points, which can only be obtained indirectly as described earlier (§4.10).

7.6. *Alti-Electrograph Results*

With the alti-electrograph (Simpson and Scrase 1937, Simpson and Robinson 1940) it has been possible to determine the sign of the field at various levels during the ascent of the balloon.

The results show that there is a positive charge in the upper part of the cloud, a negative charge in the lower part and, in some cases at least, a concentrated positive charge in a small volume in the base. The upper positive charge is centred round a region of temperature $-20°$C and the negative charge at temperatures about or just below zero, the lower positive charge being at above the freezing point.

Using the interpretation of the alti-electrograph records with Whipple and Scrase's (1936) formula, as discussed above, the conclusion was reached that the fields encountered by the balloons often reached little over $100\,\mathrm{v\,cm^{-1}}$ and did not show the increase from the ground upwards that was to be expected from the space charge of the point discharge. In view of the difference of conditions between the alti-electrograph and the point, it seems that too much reliance should not be placed on the values for the magnitudes of the fields, but this, of course, puts no doubt on the results for the signs of the fields, or on the general structure of the thundercloud deduced from these.

7.7. *Radiosonde Results*

In place of records by pole-finding paper, the point discharge current through points attached to a balloon has been measured by means of a radiosonde.

Kreielsheimer and Belin (1946) have published their results of only one such ascent, and this showed a polarity of cloud opposite to that found by the alti-electrograph.

Chapman (1950) has made similar radiosonde measurements and attempted to calibrate his apparatus with a field meter at 3000 ft. The results gave good general agreement with those of Simpson and Scrase, and the maximum field registered was only $210\,\mathrm{v\,cm^{-1}}$.

7.8. *Fields inside Clouds*

From the alti-electrograph records Simpson and Scrase (1937) and Simpson and Robinson (1940) made estimates of field strength from the width of the traces; as already discussed, such estimates may not give true values of the field. When the field exceeded the value given by their estimates as $100\,\mathrm{v\,cm^{-1}}$, further estimates could not be made since sparking occurred at the electrodes. It has frequently been stated that the alti-electrograph seldom measured fields greater than $100\,\mathrm{v\,cm^{-1}}$ whereas, in fact, a number of the ascents show fields greater than this and sufficiently large to give sparking. There are, however, several cases where the field throughout the ascent remained much smaller

than might have been expected. The matter is further complicated by the fact that it is impossible to determine exactly where the balloon travels in the cloud.

Gunn (1948) has measured fields at the surface of an aircraft in clouds, using two field machines, so as to be able to correct for the field due to charges on the aircraft itself. He found that in non-precipitating clouds the fields are less than $10\,\mathrm{v\,cm^{-1}}$, and in clouds producing steady rain less than $40\,\mathrm{v\,cm^{-1}}$. But inside active thunderclouds fields reach thousands of $\mathrm{v\,cm^{-1}}$; in one case he found a value of $-3400\,\mathrm{v\,cm^{-1}}$ just before the aircraft was struck by lightning, and he gives reasons for believing that the field inside the cloud, as distinct from that at the surface of the aircraft, was still larger than that measured.

Gunn (1950b) and others have remarked that there is a very noticeable increase in field on entering a precipitating cloud, so that a simple dipole by no means represents the electrification of the cloud.

Kuettner (1950) made observations at the Zugspitze observatory (10 000 ft), which is usually above the base of thunderclouds. His results agree very well with those from the alti-electrograph.

7.9. *Charges on Particles in Clouds*

Kuettner (1950), on the Zugspitze, found that solid precipitation particles inside clouds usually have charges of opposite sign to the field, more particularly when the temperatures are above $-5°\mathrm{c}$.

Gunn (1950a) made measurements of charges on precipitation particles within a thundercloud and also (1947) within a cloud of a weak cold front without thundery activity. In a thundercloud charges of both signs were found at most levels, and the greatest charges were found where the temperature was about $+10°\mathrm{c}$; from the particle charges alone the fields would be expected to be about 4 to 6 times greater than those actually found, so that there must exist neutralizing charges among the cloud particles or in the air. In the cold front cloud positive charges were observed at levels upwards from just below the freezing level and negative charges from the surface to well above the freezing level. Again the field to be expected from the particle charges is considerably greater than that found, so that neutralizing charges are present.

7.10. *Other Results*

Workman and Reynolds (1950b) have given several results concerning thunderstorm structure, of which the most important not so far considered are: the cloud base extends some distance below the freezing point level; electric charge separation does not occur importantly until precipitation processes are well under way; the negative charge centre lies in the precipitation zone of the cloud, but does not extend downwards in the rain below the cloud base.

The observations of the 'Thunderstorm Project' (Byers and Braham 1949) show that hail is not a necessary part of a thunderstorm, but may or may not occur. Thus theories of thunderstorm electrification which depend on hail are to be rejected.

7.11. *Lightning Results*

Although the main features of lightning have been known for some time and have been described in full by Meek and Perry (1944–5) and in more popular form by Schonland (1950), recent work on lightning has provided some fresh evidence on the electrical processes in thunderclouds.

Malan and Schonland (1951) have shown by a number of separate methods that the successive strokes of a single visual flash correspond to charge being taken to earth from successively higher regions of the cloud, rather than from regions at the same level but with horizontal separation. Mean values of the height from which a stroke starts go from 3·7 km for the first stroke up to 9·0 km for the seventh and later strokes—these being heights above the ground, itself at 1·8 km above sea level; for comparison, the cloud base is at about 2 km above the ground. However, Wormell (1953), for thunderstorms in England, found that this picture would only hold for 25% of the cases, while in the other 75% of the cases the successive strokes are concerned with charges at the same height but separated horizontally.

7.12. *The Lower Positive Charge*

Although earlier South African results gave little indication of the lower positive charge found with the alti-electrograph, more recent results (Malan 1952) show definitely that it is present, partly in the base of the cloud and partly in the air below. There is therefore now no reason to believe in any such fundamental differences between South African and English storms. Malan gives evidence for the average quantities of charge being ± 40 coulombs for the main charges and $+10$ coulombs for the lower positive charge.

Malan also considers it probable that the concentration of positive charge triggers off the lightning flash, thus approaching the views of Wichmann (see §7.13), although Malan does not postulate a separate process of separation, preferring to consider the positive charge to originate in point discharge (see §8.13).

7.13. *The Views of Wichmann*

Most writers on thunderstorm theories have taken the view that storm clouds and shower clouds are essentially similar, the difference between them being one of degree, not of kind; thus Israël (1950) and Wall (1948) consider a 'physical storm' which can be thought of as a continuously acting shower cloud in which the field never quite reaches the value for discharge. In actual storms the lightning flash is treated as an accident of secondary importance. We have followed this in §§7.4 and 7.5.

Wichmann (1952) emphatically differs from this, and puts forward the idea that there is one process acting in both shower and storm clouds to give what we have called the 'main' separation of charge, and that in storm clouds there is the additional process which gives, in a small volume, the lower positive charge with a consequent localized strong field where the lightning discharge originates. In his view the essential feature of the thunderstorm is the lightning flash, and the existence of the lower localized positive charge is essential to the lightning flash and so to the thunderstorm, as distinct from the shower.

In support of Wichmann's contention is the fact that many balloon measurements never reach fields over 1/100 of the field necessary for discharge, so that the high fields are very localized. And there is the undeniable fact that an actual storm cell has a definite life and is not continuously acting, like the 'physical storm' of Israël and Wall.

On the other hand, Wichmann's ideas lead to the conclusion that in a thunderstorm there is a separation of charge in the opposite direction to that in a shower cloud, and so one might well expect a decrease and perhaps even a

reversal of sign of the current through the cloud, whereas, in fact, Gish and Wait (1950) have shown that there is a current taking positive charge upwards through all thunderclouds. And Gunn (1948) has found large fields to cover a wide volume inside a thundercloud, showing a definite discrepancy with the balloon observations, and not favourable to Wichmann's views.

§ 8. The Origin of Cloud Charges

8.1. *The Main Charges in Clouds*

The alti-electrograph and other results all agree that the main distribution of charges in a storm cloud consists of an upper positive charge centred round a region where the temperature is about $-20°$C and a lower negative charge at a temperature somewhat below zero.

In shower clouds, which do not give lightning, the evidence is that the charge distribution is similar to that in thunderclouds, but the fields are lower.

The charges present in continuous rain clouds are of much smaller magnitudes and do not appear to have been measured directly.

In non-raining clouds there is generally found to be no noticeable charge, though on a few occasions Lecolozet (1946) found effects to be attributed to small charges in fine-weather cumulus clouds.

In storm clouds the magnitudes of the charges have been estimated to be between 10 and 100 coulombs, the external dissipating current about $\frac{1}{2}$A and the rate of separation of charge within the cloud several ampères. These results are values for a whole storm, which normally consists of several cells in various stages of development.

8.2. *Subsidiary Charges in Clouds*

In addition to the main distribution of charge there are certainly other charges present in clouds. Of these the most prominent is the positive charge in the base of the cloud found in a number of the alti-electrograph results, and existing at temperatures above the freezing point of water.

Measurements of the field at the earth's surface during storms and showers show that a simple picture of two or three centres of charge can by no means fit the results. Such observations as Simpson's (1949) 'patterns' require a much more complex arrangement of charges; though it is not certain whether the patterns are representative of a change of charge distribution with time in the cloud, or of a spatial distribution in the cloud passing over the observer, in either case a simple arrangement of charges is inadequate.

Although the first problem of the origin of cloud charges is to explain the main charges, any theory which can readily fit in the subsidiary charges is better than one which would require a separate process for these.

8.3. *Requirements of Theory*

Any theory to account for the main charging of the storm cloud must satisfy certain conditions:

(i) the process must give a positive upper and negative lower charge,

(ii) the process must give a rate of separation of charge in a storm cloud of some ampères,

(iii) the process must operate at temperatures below the freezing point of water,

(iv) the process must operate in cumulo-nimbus and perhaps in nimbo-stratus clouds, but much more effectively in the first,

(v) the process must not operate in clouds which give no precipitation.

From these requirements, and from the need for any process suggested to fit in with existing knowledge of cloud phenomena, including the initiation of precipitation, certain general results can be deduced.

It seems that the only way yet suggested in which charges of one sign can get to a different level in the cloud from charges of opposite sign is that they are attached to particles of different sizes which move, under the influence of gravity, with different speeds relative to the surrounding air. The 'particles' concerned may be ions in the air, cloud droplets, precipitation particles or perhaps small fragments of water or ice. In storm clouds there are upward air currents, and if the lighter 'particles' are carried upwards while the heavier can fall through the rising air, then there can arise a separation in space. It may be pointed out that a difference in the vertical air velocity cannot in itself give any difference in the current of separation if the same number of charged particles are present, since the separation depends on the difference of velocity relative to the air.

The connection between charge separation and precipitation suggests very strongly that it must be the precipitation particles which are the heavier particles bringing downward the negative charge to the base of the cloud. Since the main separation of charge occurs at temperatures below the freezing point, and since it is believed that nearly all precipitation originates as solid, melting only when it gets below the 0°c level, it follows that it is the precipitation particles in the solid state which are concerned in the separation of charge.

Although cumulo-nimbus clouds often give heavier rainfall than nimbo-stratus clouds, the difference in electrical activity is much greater than can be accounted for merely by greater numbers of raindrops or by larger sizes of drops. It is therefore necessary to look for some other difference between cumulo-nimbus and nimbo-stratus clouds to account for the great difference in electrical effects. We have already seen that the difference in vertical air velocity will make no difference. The other chief distinction between cumulo-nimbus and nimbo-stratus lies in the great amount of turbulence in the former, and this may well be responsible for the much greater electrical effects.

A theory to be adequate must be able to account not only qualitatively but also quantitatively for the separation of charge in clouds. A theory which predicts the wrong rate of separation of charge must be discarded, but a theory which is unable to make any prediction of rate must be regarded as not yet complete, and we must suspend judgment.

8.4. *Classification of Theories*

The theories concerned in the separation of charge in clouds can be classified in various ways. One simple way of dividing them into two classes is by the origin of the charges which are separated; in one class the charges are those on the ions existing in the clouds produced by cosmic rays and by radioactivity, and perhaps supplemented by ions produced by discharges in and near the cloud; these ions are then preferentially captured by particles, giving separation under gravitational forces. In the other class are those theories which involve a primary process of charge separation, producing fresh charges of different signs on particles of different size, so that they are then separated by gravitational action.

Wilson's theory (1929), its extension by Wall (1948) and that of Frenkel (1944) fall into the first class, and all the other important theories into the second class. Wichmann (1952) has suggested that a theory of the first class might account for the rather slow main separation of charge, but that a theory of the second class would be required for the production of the lower positive charge, which, in his view, is more rapid and leads to the lightning flash.

8.5. *Wilson's Theory*

Wilson (1929) suggested that fields in clouds could be built up by a process of selective ion capture by polarized water drops. If there exists an initial positive field, supposed by Wilson to be the fine-weather field, then the top half of a neutral drop will carry a negative charge and the lower half a positive charge. If ions of both signs move towards the drop then each half of the drop will attract ions of sign opposite to its charge, but if the drop is falling more rapidly than the downward moving ions then no ions reach the upper half of the drop and a neutral drop can only receive the upward-moving negative ions; and when the drop has received a small negative charge it still obtains more negative than positive ions until it reaches an equilibrium charge of negative sign. Thus the falling drop picks up some of the negative ions which are rising in the cloud and tending to neutralize any field in the cloud.

In itself such a process would not build up a bigger field in the cloud but would merely delay its neutralization by the ionic current. If, however, there is a vertical air current sufficient to transport upwards the ions that are moving downwards relative to the air in the field, the positive ions, and not sufficient to overcome the fall of the drops, then, relative to the earth, the negative drops fall while the excess of positive ions rise. Under such, rather limited, conditions an increase in field can occur.

Whipple and Chalmers (1944) have worked out in detail the way in which conducting spherical drops, falling according to Stokes' law, acquire charges. Wormell (1953) has extended the theory to other cases. The theory, suitably modified, can also be applied to falling ice particles (Chalmers 1947).

It remains to discuss whether the conditions envisaged by Wilson can exist in the cloud and whether the process can give sufficient separation of charge to account for the phenomena of a thundercloud. Wormell (1953) has shown that the natural rate of production of ions in the atmosphere is insufficient for Wilson's process to produce the rate of separation of charge that actually occurs. Some other sources of ions must exist, and, as Wormell points out, when the fields are high these ions cannot be small ions, as small ions travel downwards more rapidly than falling drops, and would tend to neutralize the charges on the drops.

It would seem that Wilson's process cannot be the only one operating in the cloud; some other process, even if it does not produce separation of charge itself, must provide ions for the Wilson process. And, as we have seen, Wilson's process can operate to increase the field only under rather limited conditions, which are not very likely to exist for the whole of the period of charge separation in a cloud.

It may, however, be that Wilson's process acts to assist some other process in the separation of charge and the building up of the fields. Wichmann (1952) has suggested that Wilson's process acts to produce the 'main' separation of charge, common in his view to storm and shower clouds, but that another process occurs in a storm cloud to produce the higher field needed for a lightning flash.

8.6. *Wall's Theory*

Wilson assumed that the initial field providing polarization of the falling particles was the normal fine-weather field of the earth; the process of ion capture would then cause this field to be increased. If, for any reason, there were a reversed field, then this too could be increased by the same process; such a reversed field might well be expected to occur in storm clouds, at any rate sometimes, perhaps as a result of lightning discharges or of convection of charged particles, and yet almost, if not quite, all the thunderclouds show the same main polarity. If the process of ion capture is to be the main process of charge separation, then more than the normal fine-weather field seems required to ensure that the polarity of the cloud is of the right sign.

Wall (1948) has provided a theory which overcomes this difficulty in Wilson's theory. In agreement with the requirements given above, he considers precipitating ice particles to be responsible for the charge separation, and he discusses in detail the 'asymmetry effect', which is a generalization of Wilson's ideas of selective ion capture. In place of the earth's fine-weather field to provide the initial polarization of the falling particles, Wall assumes that they are polarized by virtue of a property of ice.

Wall assumes, without as yet any direct experimental proof, that ice crystals, being polar, usually fall with their polar axes vertical and that the positive ends of the crystals are downwards as a rule. Although the exposed positive charges at the bottom of the crystal are neutralized by captured ions, fresh ice molecules condensing on to the crystal are continually providing fresh surfaces with positive charges exposed, and so the process continues.

It would seem that Wall's theory must suffer many of the disadvantages expressed above in regard to Wilson's theory, and, though it overcomes the difficulty of the definiteness of the polarity of the cloud, it requires the assumption of a process yet to be demonstrated experimentally. It is difficult to agree with Israël (1950) that Wall's is one of the most hopeful of the theories yet advanced.

8.7. *Frenkel's Theory*

Frenkel (1944) has based a theory of thunderstorm electricity on the idea that a cloud droplet, by virtue of the electrical potential difference at the water–air interface, will attract towards itself negative ions preferentially. In the vertical air current the droplets rise less rapidly than the air, and so there arises a vertical separation of charge, of the right polarity for the thundercloud.

It seems that Frenkel's theory does not satisfy a number of the requirements we have set out above. In the first place it deals with water droplets and so operates, if at all, at the wrong temperature, and, as Wall has suggested, ice crystals are likely to be polar and to attract ions of opposite signs at different places, not always ions of one sign. Then, Frenkel's process would operate whether the cloud is precipitating or not, and so we should expect to get electrical effects even with stratus and strato-cumulus clouds, and these are extremely slight if they exist at all. Frenkel's theory relies upon gravitational separation between cloud droplets and air, which is likely to be much less than between precipitation particles and air, and, further, which is likely to be less, not more, evident in very turbulent clouds.

Israël (1950) brackets Frenkel's theory with Wall's as being most worthy of consideration. The present author considers that the above arguments show that Frenkel's theory merits no further discussion.

8.8. *Ice Impact Theory*

When it was shown by the alti-electrograph results that the main separation of charge in storm clouds must occur where the temperature is below zero, and so where precipitation particles must be solid, Simpson and Scrase (1937) suggested that the process at work might be the impact of ice particles on one another, giving a negative charge to the ice fragments and a positive charge to the air. Blizzards often show high electric fields, usually positive, which would be accounted for by just the same process as Simpson and Scrase postulated.

Experimental confirmation that impact of ice on ice produces charging of the right sign is remarkably little. Pearce and Currie (1949) found that when an air blast eroded a block of snow, the larger eroded fragments acquired a negative charge and the air a positive charge; Chalmers (1952 d) found that, when ice was rubbed together, the larger fragments obtained a negative charge, in agreement with the requirements of the theory.

Sir George Simpson, in a private communication, has pointed out that 'ice friction' is not the correct description of the effect, since true friction should give no charge separation. A more correct term is 'ice impact', and possibly the operative process is the fracture of ice crystals.

It appears almost impossible to estimate whether the process suggested can give enough charge separation to produce the fields that exist in the storm cloud, and so to confirm that this is a possible process.

This theory has the advantage that it can readily be seen that the greater turbulance of the cumulo-nimbus compared with the nimbo-stratus can account for the much greater separation of charge in the first.

8.9. *Findeisen's Theory*

Findeisen and Findeisen (1943) found that when an ice surface grows by the deposition of water vapour or diminishes by evaporation there are small ice splinters which detach themselves from the surface. These ice splinters were found to be charged; in the case of growing ice the splinters carry a negative charge and the main body of ice a positive charge, while for evaporating ice the splinters are positive and the main body negative. In the presence of a strong field the sign of the splinters may be altered and then the splinters may act somewhat similarly to the Wilson theory, to increase the existing field.

Findeisen's picture may give two primary regions of charge separation, an upper one where the ice particles grow and the particles obtain a positive charge, the negative splinters dispersing, and a lower region where the ice particles evaporate, getting a negative charge, the corresponding positive charge on the splinters rising.

Israël (1950) objects to Findeisen's theory mainly because the latter also finds electrification due to riming, and this is of the wrong sign to fit with the cloud polarity. We shall discuss riming effects in a later section.

Kramer (1949) and Kumm (1951) have repeated some of Findeisen's experiments and throw doubt on Findeisen's results. In view of this discrepancy it seems of little value to discuss Findeisen's theory more fully.

8.10. *Riming Effects*

A process which occurs in storm clouds at about the right level for the production of charge is that of 'riming', in which the small supercooled water droplets in the cloud come into contact with the falling ice particles and immediately freeze on to them.

Early investigations of electrification produced by the impact of water on ice had shown a positive charge on the ice and a negative on the air or on the water droplets splashing off (Faraday 1845, Sohncke 1886, Findeisen 1940). Lueder (1951 a) pointed out that these early results were for water in the non-supercooled state, whereas in clouds it is supercooled water drops which must be concerned. Workman and Reynolds (1950 b) found that the effects produced by water falling on the ice depended largely on small amounts of contamination in the water, most normal contaminants to be found in water in the atmosphere giving a negative charge to the ice; they used a suspended hailstone.

Workman and Reynolds (1950 a, b) have developed the theory and suggest that in a suitable temperature range cloud droplets meeting falling ice particles will be frozen in part (not wholly because of latent heat requirements) and the remaining water will splash off, carrying away a charge shown by experiment to be positive, with the contaminants likely to be present in the air. At lower levels, and higher temperatures, where the ice has melted wholly or in part, a similar process of splashing carries off negative charges, and these negative charges provide the negative volume charge in the base of the cloud; if they rise and meet falling ice particles these get extra negative charges and so there is a process of 'cycling', increasing the charge separation.

Workman and Reynolds claim for their theory that it is the only one which can explain how it is that the negative charge is confined to the base of the cloud and does not extend in the rain below. They also claim that their theory is the only one to explain how it is necessary that a thundercloud always extends below the 0°c level.

Lueder (1951 a) has shown the effect to occur with natural supercooled water drops, and has also (1951 b) carried out calculations to show that the charging to be expected by this process is many times greater than that to be expected by ion capture.

While the claims mentioned above may be a little exaggerated—the theory given by Chalmers (1951) accounts for the loss of charge by raindrops below the cloud—this theory has very many merits which require it to be given very serious consideration. The only obvious drawback is a doubt as to whether the 'cycling' process described is sufficient to explain the large difference between electrical effects in cumulo-nimbus and nimbo-stratus, especially as there are still large electrical effects in cumulo-nimbus snowstorms, where the cycling process presumably cannot act.

8.11. *Other Theories*

Two other theories may be mentioned but, as they are based on hypotheses not yet experimentally verified, they need not be discussed in much detail.

Chalmers (1943) suggested that the sudden freezing of water droplets might, by destroying a free water surface, give just the opposite result to the breaking of water drops, producing fresh water surfaces; if the electrical effects are similarly opposite this would give a positive charge to the air and a negative charge to the ice particles. This would occur at just the right level to explain the upper separation of charge found by Simpson and Scrase.

Rossmann (1948) suggested electrification of ice crystals by their pyroelectric effect, and built up a thunderstorm theory on this basis.

8.12. *Discussion of Theories*

A few years ago Israël (1950) gave it as his opinion that the theories of Wall and Frenkel were those to be considered most seriously.

It is the opinion of the present author that the two theories which are best qualified to satisfy the conditions are the ice impact theory of Simpson and Scrase (1937) and the riming theory of Workman and Reynolds (1950a). With the present knowledge it does not seem possible to make any decision between the two, but both appear to be in many respects superior to those chosen by Israël.

Many writers have attempted to build up a picture of the thundercloud in which a number of different processes all played a part. While no process which can be shown to occur under the conditions existing can be neglected, it would seem to be unlikely that more than one process should be of prime importance in producing the 'main' separation of charge.

8.13. *The Lower Positive Charge*

The lower positive charge appears to be much more confined than the main charges, and it occurs where the temperature is above 0°C, so that it may well be that it is to be connected with the presence of liquid water.

Malan and Schonland (1951), followed by Malan (1952), have suggested that this lower positive charge is due to ions from point discharge at the earth's surface moving upwards in the strong updraught of the thunder cloud and becoming immobilized by attachment to cloud particles in the base of the cloud. Consideration of fields and ion paths, however, suggest that it would be only in exceptional cases that such a space charge could produce the positive fields found by the alti-electrograph; so it seems that perhaps the lower positive charge exists in its own right and not merely as a secondary effect of the main negative charge. However, this theory of the lower positive charge has the great advantage from the theoretical point of view that it does not require a distinct process of charge separation.

Simpson and Scrase (1937) suggested that this lower positive charge arose from the breaking of water drops, the process which Simpson (1909) had suggested for the main separation of charge, but which had had to be discarded both from its wrong polarity and for the reason of temperature.

Another process which would give the right sign at the right place is the charging on melting found by Dinger and Gunn (1946) to occur for ice containing entrapped air. A serious disadvantage of this process is that it would be expected to give an effect for nimbo-stratus clouds, but this is not found; however, it may be argued that a vertical separation of charge in a very widespread layer has no effect on the field below, and it might be that the negative field of the nimbo-stratus cloud is that of the negative charge from the melting process, the positive charge being that found in the rain.

Wall (1948) has made two interesting suggestions in regard to this charge. If there are negative ice particles falling and positive ions rising, at temperatures just below 0°C, with the negative charge predominating, then lower down, where the ice particles have melted, the velocity of fall of the particles will be considerably greater, and if they each carry the same charge as higher up, the charge of precipitation in unit volume of space will be much less, and so the volume charge in this region may be positive instead of negative. This idea gives

the effect occurring at just the right temperature and would not give any such effect for nimbo-stratus, where there is not the same rising current of positive ions.

Wall's other suggestion is that in the region of heaviest precipitation there is not the usual rising air current, but a falling air stream. In this, if there is the usual process of charge separation, with heavier negative and lighter positive carriers of charge, the positive charge will be brought downwards, not taken upwards, and, while the negative precipitation falls out, the positive charge remains in the lowest part of the cloud.

If Wichmann's (1952) ideas (see §7.13) are accepted, that the lower positive charge is fundamental to the production of lightning, then there must be some process which occurs only very locally and only under the special conditions of the cumulo-nimbus cloud. The breaking drop process of Simpson might meet the requirements, and Wichmann has also suggested the process of riming.

8.14. *Subsidiary Charges*

In a number of cases the field due to thunderclouds can be explained only if it can be assumed that, besides the two main charges and the lower positive charge, there are other centres of charge in the cloud. The only theory which could offer any real hope of explaining such effects might be Findeisen's if local variations altered the conditions from decrease to increase of size of ice crystals or vice versa. Otherwise these local changes of charge must be ascribed to the down-draughts that exist in thunderclouds sometimes bringing the upper positive charge down, or perhaps to effects of wind shear.

8.15. *Charge Separation in Nimbo-Stratus Clouds*

The fact that the field below a nimbo-stratus cloud is negative as a rule suggests that there is a negative charge in the base of nimbo-stratus cloud and that this might correspond to the main negative charge in the thundercloud with the same process of charge separation operating as in the thundercloud.

But Simpson's (1949) results for rain charges in fields of $+10$ to $-10\,\mathrm{v\,cm^{-1}}$ show that the rain current is positive for zero field and is zero for a positive field. Under both these conditions it would seem that there must be a positive current flowing into the earth's surface. Since the conditions are steady, there is unlikely to be any accumulation of charge going on anywhere, and hence there must be a positive charge moving downwards at all levels, a result which does not fit in with the process of separation of charge being the same as in the thundercloud.

If this conclusion is correct, then the process which operates in thunderclouds either does not operate at all in nimbo-stratus clouds or operates only with so little effect that it is outweighed by some other process, which can only be of subsidiary importance in the thundercloud.

§9. Conclusion

It will be seen from the foregoing that, although some of the problems of atmospheric electricity appear to be fairly well solved, there are others in which the solution is still apparently remote. In a subject like this, where conditions are not under control, the results of measurement are bound to be subject to many influences which cannot be assessed. And where conditions vary from one part of the world to another it may well be that observations made so far do

not form a fair sample. It is safe to predict that the time is far off when it will be possible to give a complete explanation of all the phenomena of atmospheric electricity.

ACKNOWLEDGMENTS

The author wishes here to acknowledge gratefully the support given to his work throughout many years by Professor J. E. P. Wagstaff and by the Council of the Durham Colleges, and also the help from research students working under his supervision.

REFERENCES

APPLETON, E. V., 1925, *Proc. Phys. Soc.*, **37**, 49D.
BANERJI, S. K., and LELE, S. R., 1952, *Proc. Nat. Inst. Sci., India*, **18**, 93.
BELIN, R. E., 1948, *Proc. Phys. Soc.*, **60**, 381.
BEST, A. C., 1943, *Met. Res. Ctee.*, Meteorological Research Publication 352.
BRICARD, J., 1949, *J. Geophys. Res.*, **54**, 39.
BROOKS, C. E. P., 1925, *Met. Off. Geophys. Mem.* 24.
BROWN, J. G., 1930, *Terr. Magn. Atmos. Elect.*, **35**, 1.
BRUCE, C. E. R., and GOLDE, R. H., 1942, *J. Instn. Elect. Engrs.*, **88**, Pt. II, 487.
BYERS, H. R., and BRAHAM, R. R., 1949, *The Thunderstorm* (Washington : U.S. Weather Bureau.
CHALMERS, J. A., 1939, *Quart. J. R. Met. Soc.*, **65**, 237; 1943, *Phil. Mag.*, **34**, 63; 1944, *Quart. J. R. Met. Soc.*, **70**, 121; 1946, *Ibid.*, **72**, 199; 1947, *Ibid.*, **73**, 324; 1949, *Atmospheric Electricity* (Oxford : University Press); 1951, *Quart. J. R. Met. Soc.*, **77**, 249; 1952 a, *J. Atmos. Terr. Phys.*, **2**, 155; 1952 b, *Ibid.*, **2**, 292; 1952 c, *Ibid.*, **2**, 301; 1952 d, *Ibid.*, **2**, 337; 1953 a, *Ibid.*, **3**, 223; 1953 b, *Ibid.*, **3**, 346; 1953 c, *Ibid.*, **4**, 124; 1953 d, *Ibid.*, **3**, 345; 1953 e, *Ibid.*, **4**, 84.
CHALMERS, J. A., and LITTLE, E. W. R., 1940, *Terr. Magn. Atmos. Elect.*, **45**, 451; 1947, *Ibid.*, **52**, 239.
CHALMERS, J. A., and MAPLESON, W. W., 1953, not yet published.
CHALMERS, J. A., and PASQUILL, F., 1938, *Proc. Phys. Soc.*, **50**, 1.
CHAPMAN, S., 1950, *Conference on Thunderstorm Electricity* (Chicago : University Press), p. 157.
CHIPLONKAR, M. W., 1940, *Proc. Indian Acad. Sci. Bangalore* A, **12**, 50.
DINGER, J. E., and GUNN, R., 1946, *Terr. Magn. Atmos. Elect.*, **51**, 477.
ELSTER, J., and GEITEL, H., 1888, *Met. Z.*, **5**, 95.
FARADAY, M., 1845, *Pogg. Ann.*, **60**, 342.
FEDEROV, E. K., 1951, *C. R. Acad. Sci., U.R.S.S.* (*Doklady*), **78**, 1131.
FINDEISEN, W., 1940, *Met. Z.*, **57**, 201.
FINDEISEN, W., and FINDEISEN E., 1943, *Met. Z.*, **60**, 145.
FRENKEL, Y. I., 1944, *J. Phys., U.S.S.R.*, **8**, 285.
GISH, O. H., and WAIT, G. R., 1950, *J. Geophys. Res.*, **55**, 473.
GSCHWEND, P. P., 1922, *Beilage zum Jahresber. der Kant Lehranstalt in Sarnen.*
GUNN, R., 1947, *Phys. Rev.*, **71**, 181; 1948, *J. Appl. Phys.*, **19**, 481; 1949, *J. Geophys. Res.*, **54**, 57; 1950 a, *Ibid.*, **55**, 171; 1950 b, *Conference on Thunderstorm Electricity* (Chicago : University Press), p. 162.
HARPER, W. R., 1934, *Phil. Mag.*, **18**, 97; 1935, *Ibid.*, **20**, 740.
HESS, V. F., and O'DONNELL, G. A., 1951, *J. Geophys. Res.*, **56**, 557.
HOGG, A. R., 1934, *Gerlands Beitr. Geophys.*, **41**, 1; 1939, *Mem. Comm. Solar Obs., Canberra.*
HUTCHINSON, W. C. A., 1951, *Quart. J. R. Met. Soc.*, **77**, 627.
HUTCHINSON, W. C. A., and CHALMERS, J. A., 1951, *Quart. J. R. Met. Soc.*, **77**, 85.
ISRAËL, H., 1941, *Gerlands Beitr. Geophys.*, **57**, 261; 1948, *Met. Runds.*, **1**, 344; 1950, *Das Gewitter* (*Probleme der kosmischen Physik*, **25**), p. 93.
ISRAËL, H., and KASEMIR, H. W., 1952, *Arch. Met. Geophys. Biokl.*, **5A**, 71.
ISRAËL, H., KASEMIR, H. W., and WIENERT, K., 1951, *Arch. Met. Geophys. Biokl.*, **3A**, 357.
KASEMIR, H. W., 1950, *Arch. Met. Geophys. Biokl.*, **3A**, 84; 1952, *J. Atmos. Terr. Phys.*, **2**, 1.
KRAMER, C., 1949, *Mededelingen en Verhandelingen*, **54A**, No. 102.

KREIELSHEIMER, K., and BELIN, R. E., 1946, *Nature, Lond.*, **157**, 227.
KUETTNER, J., 1950, *Conference on Thunderstorm Electricity* (Chicago : University Press), p. 99.
KUMM, A., 1951, *Arch. Met. Geophys. Biokl.*, **3A**, 382.
LECOLOZET, R., 1946, *La Météorologie*, 429.
LUEDER, H., 1943, *Met. Z.*, **60**, 340; 1951 a, *Z. angew. Phys.*, **3**, 247; 1951 b, *Ibid.*, **3**, 288.
LUTZ, C. W., 1941, *Gerlands Beitr. Geophys.*, **47**, 317.
MALAN, D. J., 1952, *Ann. Geophys., Paris*, **8**, 385.
MALAN, D. J., and SCHONLAND, B. F. J., 1951, *Proc. Roy. Soc.* A, **209**, 158.
MAPLESON, W. W., 1953, not yet published.
MATHIAS, A., 1926, *Elektrizitätswirtschaft*, **25**, 297.
MAUCHLY, S. J., 1923, *Terr. Magn. Atmos. Elect.*, **28**, 61.
MEEK, J. M., and PERRY, F. R., 1944–5, *Rep. Prog. Phys.*, **10** (London : Physical Society), p. 314.
MÜHLEISEN, R., 1951, *Z. Naturforsch.*, **6A**, 667.
NOLAN, J. J., 1950, *Arch. Met., Wien*, **3A**, 17.
NOLAN, J. J., BOYLAN, R. K., and DE SACHY, G. P., 1925, *Proc. Roy. Irish Acad.* A, **37**, 5.
NOLAN, J. J., and GUERRINI, V. H., 1935, *Proc. Roy. Irish Acad.* A, **43**, 5.
NOLAN, J. J., and NOLAN, P. J., 1937, *Proc. Roy. Irish Acad.* A, **43**, 79.
NOLAN, J. J., NOLAN, P. J., and GORMLEY, P. G., 1938, *Proc. Roy. Irish Acad.* A, **45**, 47.
NOLAN, J. J., and DE SACHY, G. P., 1927, *Proc. Roy. Irish Acad.* A, **37**, 71.
NOLAN, P. J., and DOHERTY, D. J., 1950, *Proc. Roy. Irish Acad.* A, **53**, 163.
NOLAN, P. J., and KENNAN, E. L., 1949, *Proc. Roy. Irish Acad.* A, **52**, 17.
NOLAN, P. J., and POLLAK, L. W., 1946, *Proc. Roy. Irish Acad.* A, **51**, 9.
NOLAN, P. J., and TEDDE, P. G., 1950, *Arch. Met. Geophys. Biokl.*, **2A**, 374.
O'DONNELL, G. A., 1952, *J. Atmos. Terr. Phys.*, **2**, 201.
PEARCE, D. C., and CURRIE, B. W., 1949, *Canad. J. Res.* A, **27**, 1.
RANGS, 1942, S.I.G.E.S.C.O. Report IIB/6022 "German Research in Wartime": Ministry of Supply).
ROSSMANN, F., 1948, *Met. Rundsch.*, **1**, 193.
RUSSELTVEDT, N., 1925, *Beih. Jahrb. Norw. Met. Inst.*
SCHONLAND, B. F. J., 1928, *Proc. Roy. Soc.* A, **118**, 252; 1950, *The Flight of Thunderbolts* (Oxford : University Press).
SCHONLAND, B. F. J., and ALLIBONE, T. E., 1931, *Nature, Lond.*, **128**, 784.
SCRASE, F. J., 1933, *Met. Off. Geophys. Mem.* 58; 1935, *Ibid.*, 67; 1938, *Ibid.*, 75.
SIMPSON, G. C., 1909, *Phil. Trans. Roy. Soc.* A, **209**, 379; 1910, *Phil. Mag.*, **19**, 715; 1926, *Proc. Roy. Soc.* A, **111**, 56; 1949, *Met. Off. Geophys. Mem.* 84.
SIMPSON, G. C., and ROBINSON, G. D., 1940, *Proc. Roy. Soc.* A, **177**, 281.
SIMPSON, G. C., and SCRASE, F. J., 1937, *Proc. Roy. Soc.* A, **161**, 309.
SMITH, L. G., 1951, *Dissertation*, University of Cambridge.
SOHNCKE, L., 1886, *Ann. Phys., Lpz.*, **28**, 551.
TORRESON, O. W., and WAIT, G. R., 1934, *Terr. Magn. Atmos. Elect.*, **39**, 47.
TRICHEL, G. W., 1938, *Phys. Rev.*, **54**, 1078.
WADDEL, R. C., DRUTOWSKI, R. C., and BLATT, W. N., 1946, *Proc. Inst. Radio Engrs. and Waves and Electrons*, **34**, 161.
WALL, E., 1948, *Wetter Klima*, **1**, 7, 65, 193, 321.
WATSON, R. E., 1929, *Met. Off. Geophys. Mem.* 45.
WHIPPLE, F. J. W., 1929, *Quart. J. R. Met. Soc.*, **55**, 351; 1933, *Proc. Phys. Soc.*, **45**, 367.
WHIPPLE, F. J. W., and CHALMERS, J. A., 1944, *Quart. J. R. Met. Soc.*, **70**, 103.
WHIPPLE, F. J. W., and SCRASE, F. J., 1936, *Met. Off. Geophys. Mem.* 68.
WHITLOCK, W. S., 1953, not yet published.
WICHMANN, H., 1951, *Arch. Met. Geophys. Biokl.*, **3A**, 290; 1952, *Ibid.*, **5A**, 187.
WILSON, C. T. R., 1916, *Proc. Roy. Soc.* A, **92**, 555; 1920, *Phil. Trans. Roy. Soc.* A, **221**, 73; 1925, *Proc. Phys. Soc.*, **37**, 32D; 1929, *J. Franklin Inst.*, **208**, 1.
WORKMAN, E. J., and HOLZER, R. E., 1939, *Rev. Sci. Instrum.*, **10**, 160.
WORKMAN, E. J., and REYNOLDS, S. E., 1950 a, *Phys. Rev.*, **78**, 254; 1950 b, *Conference on Thunderstorm Electricity* (Chicago : University Press), p. 31.
WORMELL, T. W., 1927, *Proc. Roy. Soc.* A, **115**, 443; 1930, *Ibid.*, 127, **567**; 1953, *Quart. J. R. Met. Soc.*, **79**, 3.

THE PHYSICAL CONDITION OF THE SOLAR CORONA

By C. W. ALLEN
University of London Observatory, London

CONTENTS

		PAGE
§ 1.	Introduction	135
§ 2.	Observations and description of the corona	136
§ 3.	Density distribution	137
§ 4.	Temperature	140
§ 5.	Ionization	143
§ 6.	Excitation	144
§ 7.	Chemical composition	146
§ 8.	Variations with the sunspot cycle	147
§ 9.	Radiation and energy balance	149
§ 10.	Dynamical structure of the corona	150
	Acknowledgments	152
	References	152

Abstract. The observations from which the physical conditions of the corona may be derived have been reviewed. The density distribution obtained from the light scattering of electrons is a straight mean. However, the corona is an irregular object and some attempt is made to determine the amount and the effect of the irregularity. A comparison is drawn between estimates of temperature by line breadth, density gradient, ionization and radio emission. Recent evaluations are given of the constants in the ionization formula. The excitations of coronal lines by collisions and by solar radiation are discussed quantitatively and lead to an excitation collision cross section of approximately 2×10^{-17} cm^2. The estimate of the metal to hydrogen ratio in the corona is rather higher than in other astrophysical sources.

The main sunspot cycle changes are (*a*) a variation in the intensity and distribution of streamers and (*b*) a variation of a 1·8 factor in electron density with a density minimum about a year after sunspot minimum. Temperature variations are only about 15% but are fairly regular.

Present knowledge of the source of heating in the corona is too indefinite to allow quantitative estimates of the heat gained by the corona, but the heat lost by conduction and ultra-violet radiation is of the order of 2×10^4 erg cm^{-2} sec^{-1}. The streamers and geomagnetic storms give evidence of a flow of material outward through the corona. This appears to be 1000 times greater than the thermal evaporation and calls for outward forces in the corona that are not at present understood.

§ 1. Introduction

Until about 1942 the solar corona was a challenge to astronomers and physicists. Its spectrum was known but not understood. The bright continuum observed close to the sun's limb was thought to be due to electron scatter, but the absence of Fraunhofer lines was unexplained. Fraunhofer lines were indeed detected weakly in the outer corona but were apparently not due to the same scattering source as the inner continuum. This left the absorption lines also unexplained. Finally there were several bright emission lines which

were not identified and were usually called 'coronium' lines. These lines were eventually identified by Edlén (1942) who showed them to be due to the elements A, Ca, Fe, Ni in very high states of ionization. The high temperature required to explain the ionization also explained (*a*) the absence of Fraunhofer lines in the bright near continuum, (*b*) the absence of the more familiar emission lines of H and the metals, (*c*) the breadth of the coronal emission lines, (*d*) the extended size of the corona, and (ϵ) the new (1946) radio measurements of quiet sun intensity.

As a result of these successes the corona is now as well understood physically as most astronomical objects. In ten years it has changed from an enigma to an astrophysically useful body in which one can study the activities of atoms at high temperatures. The conditions of temperature and pressure may be derived fairly directly from observations and could attain a good level of accuracy if it were not for the irregularity and variability of the corona itself. From the temperature and pressure distribution one may estimate the emission of thermal radio waves and the conditions of excitation and ionization. One finds fairly consistent agreement with observation. On the other hand the energy balance, structural forms and dynamic activity of the corona are not well understood.

The problems of the corona have been reviewed from several points of view by Unsöld (1938), Waldmeier (1941), Bugoslavskaya (1950), Siedentopf (1950), Mitchell (1951), Shklovskij (1951) and Woolley and Stibbs (1953). Without doubt a comprehensive analysis will be given by Waldmeier (1953+). However, progress with the coronal problems is closely bound up with our knowledge of its physical condition, which is therefore the subject of this review.

§2. Observations and Description of the Corona

For many years astronomers have used the opportunities provided by the solar eclipses to observe the colour, spectrum, brightness distribution, polarization and appearance of the corona. Many of these observations may now be made by coronagraph instead of by eclipse. The coronagraph method allows one to attain higher resolution and much better observational continuity, but it is available only for the bright inner corona.

The advent of radio astronomy has provided powerful methods of studying the corona. Radio observations may be made without eclipses, and by frequency selection one may obtain information for both the high and low corona. Distribution of radio intensity across the sun's disc and out into the corona may be determined by radio interferometry or by use of eclipses.

Geomagnetic and ionospheric observations give evidence of the variable particle and ultra-violet radiations that come from the sun and therefore through the corona. These must be taken into consideration in assembling our ideas of the physical corona.

The corona is found to have a spectral distribution very similar to that of the sun (Ludendorff 1925, Grotrian 1931). The emission spectrum lines contribute only about 1% of the total light and therefore do not influence the measurements of colour, polarization and brightness. The brightness decreases rapidly from the limb outward.

There is no well-defined outer boundary to the corona, and the limit to which the corona can be detected depends rather fortuitously on the extent of the streamers (coronal rays). These can normally be seen to about four radii from the sun's centre but have been traced as far as twelve radii (Laffineur *et al.* 1952). On the

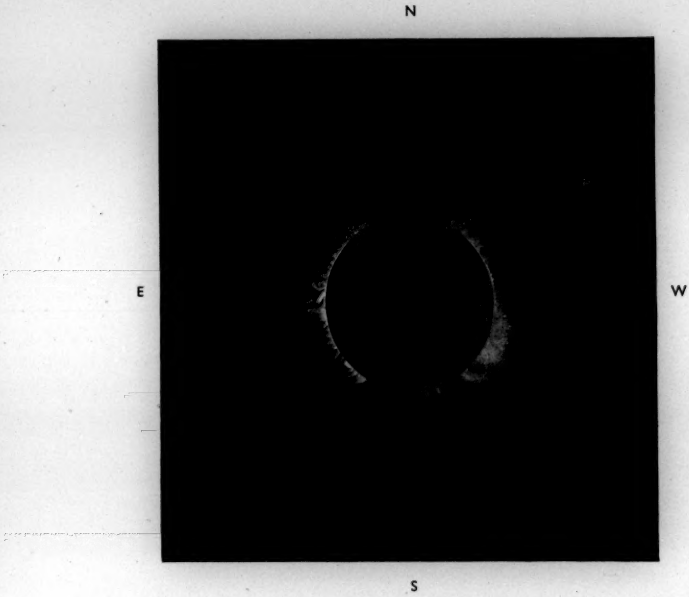


Fig. 1. Drawing of the corona of 1901 May 18 (from *Mem. R. Astr. Soc.*, 1927, **64**, Appendix, Plate 11). *Reproduced by courtesy of the Royal Astronomical Society.*

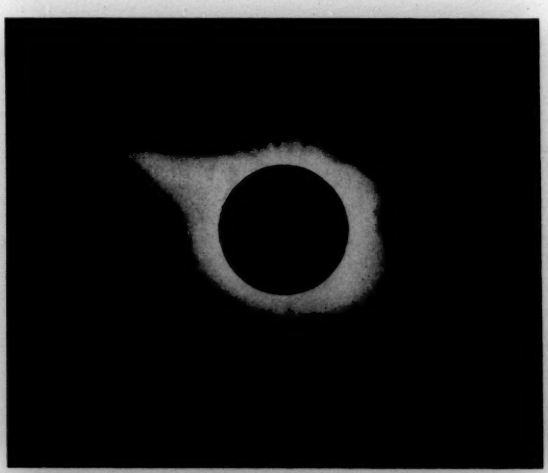


Fig. 2. Photograph of the corona of 1932 August 31 (photograph by P. A. McNally, *reproduced with permission from* '*Astronomy*' by R. H. Baker (New York: Van Nostrand, 1938), p. 296).

other hand the lower boundary between the chromosphere and the corona is rather sharp. The physics of this boundary is a difficult problem which appears again in connection with the existence of prominences.

It is evident from polarization measurements that the corona is essentially an extended solar atmosphere which scatters sunlight and also emits certain spectrum lines. The Fraunhofer absorption lines are completely obliterated in the process of scattering the light. However, superimposed on the corona proper is a halo of light diffracted by interplanetary (zodiacal light) particles. This component of the radiation does contain the Fraunhofer lines and therefore was called by Grotrian (1934) the F corona to distinguish it from the purely continuous K corona. The F corona contributes a greater proportion of the total brightness as the distance from the sun increases. For studies of the coronal physical condition it is necessary to ensure that brightness observations are suitably corrected for the F corona.

The appearance of the corona shows that it cannot be regarded as a quiet atmosphere in equilibrium under gravity alone. The structural features can best be demonstrated by sketches such as shown in fig. 1 (Plate) for the eclipse of 1901 May 18, and photographs as in fig. 2 (Plate) for the eclipse of 1932 August 31. The appearance differs markedly from one eclipse to another and is far more complicated at time of sunspot maximum. The illustration in fig. 1 is a minimum eclipse which, however, shows the main features as follows:

(*a*) Polar plumes are well developed at the N and S poles. At sunspot maximum these plumes are confused by other features.

(*b*) A set of equatorial plumes can be detected on the E side. These radiate from active areas and usually—but not always—from sunspots. At sunspot maximum this type of plume may be found at any solar latitude.

(*c*) Smooth streamers can be seen emerging from most of the sun's surface in fig. 1, and a remarkable one is shown in fig. 2. Further from the sun's surface they tend to gather into rather straight rays. The base of the streamers often encloses a mitre-shaped area containing arches and prominences.

(*d*) A set of arches surmounting a prominence may be seen in the south-east quadrant of fig. 1.

The polar plumes are the only evidence that the sun possesses a general dipole magnetic field which penetrates into the corona. The plumes probably consist of ionized gas constrained to move along lines of force of the dipole field. It is tempting to conclude that the equatorial plumes also represent the lines of magnetic force due to sunspots (or magnetically active areas). The difficulty with the general application of this idea is that the plumes appear to run parallel to the streamer edges and would suggest that the lines of the magnetic field follow the streamer edges into the distant outer corona. Magnetic fields of this shape are not possible. Coronal observations give no indications of the magnitudes of the magnetic fields at the poles, but fields associated with equatorial plumes may be derived from sunspot observations (Giovanelli 1947, Grotrian 1948).

§3. Density Distribution

The distribution of density in the solar corona may be determined rather directly from photometry of coronal brightness. It is now evident that the coronal continuum is caused by sunlight scattered by an electron atmosphere. The chief arguments for this are that the colour of the corona is the same as the sun, and that

no particles other than electrons could have a velocity large enough to obliterate the Fraunhofer lines.

The light scattering coefficient of electrons is known from Thomson's formula

$$\sigma_e = (8\pi/3)(e^2/mc^2)^2 = 0{\cdot}665 \times 10^{-24}\ \text{cm}^2. \qquad \ldots\ldots(1)$$

The radiation emitted by the sun's surface is also accurately known. Hence, if one were to assume a particular distribution of electrons in the coronal region, one could compute the projected distribution of light as seen from the earth. The coronal electron density problem is the reverse one of converting a measured projected distribution into a distribution in space. Analyses have been made by Minnaert (1930), Baumbach (1937), and Van de Hulst (1950). It has been necessary to make some assumptions on the regularity and symmetry of the electron distribution. Baumbach (1937) used spherical symmetry and a density distribution that could be expressed in three power terms. His result, which has been widely used, gives the mean electron density

$$N_e = (0{\cdot}036r^{-1{\cdot}5} + 1{\cdot}55r^{-6} + 2{\cdot}99r^{-16}) \times 10^8 \text{ electrons cm}^{-3}$$

where r is the radial distance from the sun's centre in terms of the sun's radius $R_\odot$. However, the $0{\cdot}036r^{-1{\cdot}5}$ term in this expression is due to the inclusion of the F corona in Baumbach's analysis (Allen 1947), and this term should be omitted. The more recent analysis by Van de Hulst (1950) extracts the F corona, allows for sunspot cycle variability, and at sunspot minimum also allows for the ellipticity of the corona. Van de Hulst's results are the best available and are quoted in table 1.

Table 1. Electron Densities in the Solar Corona (from Van de Hulst)

	N_e (units of 10^8 electrons cm^{-3})			
r	Spot	Spot minimum		x
(units of $R_\odot$)	maximum	Equator	Pole	
1·0	403	227	174	1·0
1·03	316	178	127	
1·06	235	132	87·2	
1·1	160	90·0	53·2	
1·2	70·8	39·8	16·3	1·03
1·3	37·6	21·2	5·98	
1·5	14·8	8·30	1·41	1·06
1·7	7·11	4·00	0·542	1·11
2·0	2·81	1·58	0·196	1·22
2·6	0·665	0·374	0·040	1·73
3·0	0·313	0·176	0·017	2·2
4·0	0·090	0·050	0·004	4
5	0·044	0·025		6
6	0·029	0·016		9

The atom density may be readily determined when the electron density is known. Since the coronal atmosphere is dominated by completely ionized hydrogen the atom density N_a will be approximately the same as N_e. One could allow for the abundance of helium and heavier elements if necessary. However, it is sufficient to put $N_a = N_e$ and consider the atoms to have a mean mass of 2×10^{-24} g. The mean molecular weight for all coronal particles is taken as 0·67.

The densities found from the brightness distribution are mean densities averaged in two ways which must be taken into consideration if one is to use the values in table 1 :

(*a*) The data used for the analysis have been averaged from the observations at many eclipses and the corona at any one might differ from this average. It would appear from the data quoted by Baumbach (1937), Dyson and Woolley (1937), and Van de Hulst (1950) that the dispersion in either the total coronal brightness or the brightness at any specified radial distance (for a particular phase of the sunspot cycle) is $\pm 0{\cdot}3$ mag, or about $\pm 25\%$. The dispersion in smoothed electron density should not be much greater than this, and it is suggested that $\pm 30\%$ on to the values quoted in table 1 should give a safe indication of electron density for $r<2{\cdot}5$. The dispersion would probably increase for $r>2{\cdot}5$.

(*b*) The electron densities so found refer to the mean densities over volumes that are large enough to include several of the corona's structural features. Since the amount of scattered light is proportional to density it is clear that the structural features represent real irregularities in the electron and atom densities. The mean densities of table 1, found from electron scatter methods, will represent straight means over large volumes. For many purposes this is not the value required. Suppose that the corona were built of dense features occupying $1/x$ of the volume and separated by very rarefied regions. Then the density of the features would be $x\overline{N}_e$. For physical studies connected with emission spectra it is the density $x\overline{N}_e$ that will be required, while for other purposes it will be the mean square density $x^{1/2}\overline{N}_e$. We could define x more generally as $x=\overline{N_e^2}/(\overline{N}_e)^2$. There is no accurate method of estimating x at present but rough values may be obtained. Following Van de Hulst (1950) we determine Δr, the mean difference in radius of isophotes for intense regions as compared with fainter regions ($\pm\frac{1}{2}\Delta r$ is the mean departure above and below the smoothed elliptical isophote). Isophotes by Bugoslavskaya (1950), O'Brien *et al.* (1939), and Barabashev (1939) lead to a result that may be expressed $\Delta r\simeq 0{\cdot}1\,(r-1)$. Eliminating the F corona with the help of Van de Hulst's data we obtain the results given in table 2. First Δr is converted into the ratio of the bright to the faint areas of the K (electron scatter) corona. It is then assumed that the line of sight depth of the features is one-third of the total line of

Table 2. Estimation of x

r	1·0	1·5	2·0	2·5	3	4	>4
Δr	0·0	0·05	0·1	0·15	0·2	0·3	
K corona brightness ratio (bright/faint)	1·1	1·2	1·5	2·2	4	∞	∞
Brightness ratio (feature/background)	1·3	1·6	2·5	4·6	10	∞	∞
x	1·02	1·06	1·2	1·6	2·2	>3	$\frac{1}{3}r^2$

sight depth through the corona, and x is computed from $\overline{N_e^2}/(\overline{N}_e)^2$. This would give $x=3$ for all high values of r, whereas it should certainly increase with r. The visible corona beyond $r=4$ consists of streamers only, and these do not appear to increase in cross section as r increases. Hence $x\propto r^2$ in the outer corona, and to fit on to the value at $r=3$ we adopt $x=\frac{1}{3}r^2$ for $r>4$. The values of x obtained are given in both table 1 and table 2. Shklovskij and Pickelner (1951) have found $x=2$ or 3 by radio methods.

The estimation of density from photometry is so direct that below the level $r=2{\cdot}0$ the density values so obtained must be preferred to those obtained by other methods. For heights above $r=2{\cdot}5$ however uncertainties increase : the corona

is more variable, the x factor is much greater, and only a small proportion of the radiation received actually comes from electron scatter. It is therefore for the higher levels that we seek information from other methods.

Evidence of density in the high corona comes from measurements of the radio occultation of the Taurus discrete radio source. It was found by Machin and Smith (1952) that this source began to decrease in apparent intensity at $r = 10$. This is not interpreted as an absorption of radio waves (which would demand much too high a coronal density), but as a refractive diffusion of the source which renders it less detectable by the usual radio-interference method. Machin and Smith have estimated the diameter of the diffused image to be as table 3.

Table 3

r	Diameter of Image		Deduced greatest N_e (electrons cm^{-3})		xN_e (0·75 × sun-spot max)
	38 Mc/s	81·5 Mc/s	38 Mc/s	81·5 Mc/s	
4·5	45′	25′	23×10^4	60×10^4	22×10^4
10	15′	8′	8×10^4	20×10^4	

The maximum angular refraction in radians will be approximately $e^2 N_e / 2\pi m f^2$, which is the difference between the refractive index and unity where f is the frequency. Equating this to the radii of the diffused images we obtain the maximum density value of N_e. These values agree reasonably well with $x\overline{N_e}$ where $\overline{N_e}$ is taken midway between sunspot maximum and sunspot minimum from table 1.

Bell (1953, unpublished, reported by P. A. O'Brien) has studied the recent radio measurements by O'Brien (1953) and finds he cannot reconcile the results with the Van de Hulst electron densities of table 1. From an analysis similar to that of Smerd (1950) he finds the electron densities of table 4. The analysis does not allow for the irregularities of the corona, and since N_e^2 is the factor appearing in Smerd's formulation the results might be regarded approximately as an estimation of $(\overline{N_e^2})^{1/2}$ and comparable with $x^{1/2}N_e$ of table 1. The comparison is shown in table 4 where we use $\overline{N_e} = 0{\cdot}75 \times (\overline{N_e}$ for sunspot maximum). The agreement is satisfactory but suggests that the values of x should be increased.

Table 4

r	1·2	1·5	1·8	2	3	4
$(\overline{N_e^2})^{1/2}$ (10^6 electrons cm^{-3}) (Bell)	68	14	4·6	3·1	0·6	0·2
$x^{1/2}\overline{N_e}$ (10^6 electrons cm^{-3}) (table 1)	53	11	4·1	2·3	0·35	0·14

§4. Temperature

The various reasons for the belief in a high coronal temperature that were enumerated in §1 all indicate a kinetic temperature in the neighbourhood of a million degrees. The radiation density temperature, which depends on the total solar radiation, is about 4000°K in the lower corona and decreases at higher radial distances r in accordance with $T = T_E(2r)^{-1/2}$ where T_E is the sun's effective temperature 5780°K. The excitation temperature will lie between the radiation temperature and the kinetic temperature, its value being dependent on the relative effectiveness of radiation and collisions in producing the excitations.

The most direct estimates of temperature are from breadths of coronal emission lines. The measurements by Lyot (1937), Waldmeier (1945), and Dollfus (1953) give values near $2{\cdot}0 \times 10^6$ deg. K. Some of Waldmeier's measurements (1947) lead to greater temperatures, up to 6×10^6 deg. K in active areas. Waldmeier (1946, 1947) also finds the line widths to decrease with height up to $r \simeq 1{\cdot}4$ at which level the line breadth temperature must be less than 10^6 deg. K. No line breadth measurements can be made in the faint outer corona.

The absence of the familiar emission lines of hydrogen and the metals in the coronal spectrum indicates that the temperature is at least $0{\cdot}7 \times 10^6$ deg. K (Goldberg and Menzel 1948).

A careful comparison of wide depressions in the coronal continuum in the neighbourhood of the solar H and K lines of Ca II could yield a measurement of kinetic electron temperature. Grotrian's observations (1931) led to a temperature of about a million degrees. However, it has not been possible to recognize the H and K depressions in the corona continuum with any certainty (Shajn 1947), and the chief result of the observations is to conclude that the temperature is very high.

Table 5. Estimates of Coronal Temperature (in 10^6 deg. K)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1·0	2·0	1·3	1·3	1·0	0·7	0·8	0·9
1·2	2	1·5	1·5	1·1	0·9	0·8	1·1
1·5	<1	1·6	1·6	1·2	0·8	0·5	0·9
2·0		1·3	1·3	1·2	0·8	0·22	0·7
2·5		1·1	1·1	1·0		0·12	0·4
3		1·0	1·0	0·9		0·07	
4		1·0	1·0	0·8		0·03	
6		1·0	1·0	0·7			

(1) r (in units of $R_\odot$); (2) from line breadth; (3)–(5) from density gradient: (3) sunspot maximum, (4) sunspot minimum, equator, (5) sunspot minimum, pole; (6) from ionization; (7) from radio emission (Bell); (8) mean.

In an atmosphere in gravitational equilibrium one may estimate the temperature from the pressure scale height H (height increase for one exponential fall of pressure), the mean molecular weight μ, and gravity g_0/r^2. The relation is (Van de Hulst 1950)

$$\frac{1}{T} = -\frac{kr^2}{g_0 \mu m_H R_\odot}\left(\frac{d \ln N_e}{dr} + \frac{d \ln T}{dr}\right)$$

$$= \frac{kr^2}{g_0 \mu m_H H} \qquad \ldots\ldots(2)$$

where μ is the mean molecular weight, m_H the mass of the hydrogen atom, k Boltzmann constant, $R_\odot$ the sun's radius, and r the radial distance in terms of $R_\odot$. The term representing change of temperature $d \ln T/dr$ can be neglected in comparison with the term $d \ln N_e/dr$, representing change of electron density. With $g_0 = 2{\cdot}74 \times 10^4 \,\mathrm{cm\,sec^{-2}}$, $\mu = 0{\cdot}67$, and the density distribution N_e of table 1, one obtains the temperatures of table 5, columns (3), (4) and (5). The results found in this way are of the right order of magnitude, but there are several reasons to suspect the accurate application of eqn. (2). The structural features indicate

that there are other forces that cannot be neglected in comparison with gravity. Furthermore, the question of whether one should use $\overline{N}_e$ or $x\overline{N}_e$ in eqn. (2) is dependent on whether the level considered is above or below the exosphere (level above which atoms suffer no collisions in moving to or from infinity). If we consider the coronal material to be forced into streamers by constraints which do not affect gravity, then below the exosphere we should use $x\overline{N}_e$ (by application of the Boltzmann equation), and above the exosphere we should use $\overline{N}_e/r^2$ (since constraints would make no difference to the total number of atoms at a particular level, and these would be dispersed in proportion to r^{-2}). From § 10 we find the exosphere to be at $r \simeq 3$. These complications are not very severe in the corona below $r = 2$, where the derived values of T must be fairly sound. However, the formula can give no reliable indication of the run of temperature in the outer corona.

In order to derive the temperature from the degree of ionization one needs to know the ionization cross-sections and the recombination coefficients for the ions concerned. The question is treated in § 5. The derived temperature is not very sensitive to the values adopted for these coefficients nor to the spectrum line intensities from which the calculation is made. Elwert (1952) found that the entire observed range of relative intensities of the lines 5303 of Fe XIV and 6374 of Fe X could be explained by temperatures in the narrow range 0·6 to 0·8 million degrees. Waldmeier's temperature observations (1952) show the pole to be about 0·2 million degrees lower than the equator, and the base level $r = 1{\cdot}0$ to be 0·2 million degrees lower than the mid-level $r = 1{\cdot}2$. The mean ionization temperature for the low corona is about $0{\cdot}8 \times 10^6$ deg. K. The occasional appearance of the lines 5445, 5694 may be interpreted (Waldmeier 1951) as an indication of localized high temperature, although the identification of these lines is still in doubt (Garstang 1952). With regard to the higher corona (up to $r = 2{\cdot}2$), it has been found by Allen (1946) that the relative intensities of the green and red lines remain nearly constant and therefore give no indication of a fall of temperature. Up to this level there are enough ionizing collisions for the application of the ionization formulae of § 5.

The radio emission from the sun in the quietest conditions is of thermal origin and gives another means of determining the temperature. Frequencies above 200 Mc/s are emitted from both the corona and chromosphere and give useful information of the chromosphere–corona transition zone (Piddington 1950, Hagen 1951, Woolley and Allen 1950) without providing much information on the corona itself. The lower frequencies tend to be disturbed by solar activity but can be interpreted in terms of coronal temperature. The Woolley and Allen (1950) value is $0{\cdot}7 \times 10^6$ deg. K, and Waldmeier (1948) gives $1{\cdot}4 \times 10^6$ deg. K. More recent observations permit a much better allowance for the spread of radiation beyond the sun's limb, and Bell's preliminary analysis (1953, unpublished, reported by P. A. O'Brien) of O'Brien's (1953) observations is given in table 5. It can be seen that the fall of temperature in the outer corona is much more rapid than that found by other methods.

The last column in table 5 gives the author's opinion of the best values to adopt at present. The line breadth observations are larger and give evidence of some high turbulent velocities in the lower corona, the scale of the turbulence being smaller than 10^4 km and therefore not showing individual Doppler shifts (Waldmeier 1947).

§5. Ionization

The elements revealed by coronal emission lines are in the ionization stages X to XVI, and the corresponding lower ionization potentials vary from 233 volts for Fe X to 655 volts for Ca XIII (omitting Ca XV and A XIV as doubtful). The Saha ionization equation (which would lead to temperatures of about 100 000°K) is not applicable because the kinetic and radiation-density temperatures differ enormously. The degree of ionization depends on the four processes by which ions are formed and disappear. These are :

Ion formation	Ion disappearance
(*a*) Radiative ionization.	(*c*) Radiative recombination.
(*b*) Electronic collision.	(*d*) Three-body collisions.

The relative effectiveness of these four processes in the special conditions of the corona has been considered by Biermann (1947), Woolley and Allen (1948), Miyamoto (1949), and Elwert (1952) who show that only (*b*) and (*c*) need be taken into account. This leads to the simple ionization formula :

$$\frac{N_2}{N_1} = \frac{S_{12}}{\alpha_{21}} \qquad \ldots\ldots(3)$$

where N_1 and N_2 are the number densities of ions in the lower and upper stages of ionization, and S_{12} and α_{21} are the coefficients of ionization and recombination. S_{12} and α_{21} are defined through the relations :

$$\text{Number of ionizations by electronic collisions} = N_1 N_e S_{12}$$
$$\text{Number of recombinations} = N_2 N_e \alpha_{21}.$$

The electron density N_e does not enter the ionization formula.

A recent study of S_{12} and α_{21} has been made by Elwert (1952). His result for S_{12} may be written

$$S_{12} = 2{\cdot}20 \times 10^{-8} \left(\frac{\chi_{12}}{kT}\right)^{-1/2} \left(\frac{\chi_{12}}{\chi_H}\right)^{-3/2} \exp(-\chi_{12}/kT) c_2 f \zeta G \qquad \ldots\ldots(4)$$

where the constant $2{\cdot}20 \times 10^{-8}\,\text{cm}^3\,\text{sec}^{-1} = 3\,\sigma_e c/4\pi^{1/2}\alpha_f{}^3$ (with α_f the fine structure constant), χ_{12} is the ionization potential between lower and upper stage of ionization, T is the electron temperature, and c_2, f, ζ, and G are all of order unity; $c_2 \simeq 2{\cdot}0$ is a parameter in the curve connecting the ionization cross section with electron energy, $f \simeq 0{\cdot}8$ is a numerical factor representing the error of the theory, $\zeta = 3$ to 7 is the number of electrons in the outer shell of the lower ion, and $G \simeq 0{\cdot}8$ is related to the Maxwellian electron velocity distribution. Choosing $\zeta = 3$ to obtain fair agreement with experiment (Bates *et al.* 1950) we have the result

$$S_{12} = 4 \times 10^{-8}\, T^{1/2} \chi_{12}{}^{-2} \exp(-11\,600\,\chi_{12}/T)\,\text{cm}^3\,\text{sec}^{-1} \qquad \ldots\ldots(5)$$

where χ_{12} is now expressed in volts and T in °K.

For the recombination coefficient α_{21} Elwert gives

$$\alpha_{21} = 5{\cdot}2 \times 10^{-14} \left(\frac{\chi_H}{kT}\right)^{1/2} \left(\frac{\chi_{12}}{\chi_H}\right) n f G g \qquad \ldots\ldots(6)$$

where the constant $5{\cdot}2 \times 10^{-14}\,\text{cm}^3\,\text{sec}^{-1} = 8\sigma_e c/(3\pi)^{1/2}$, and the last four factors are again of the order unity. $n = 3$ is the quantum number of the highest ground-state shell, f and G are approximately the same as for the S_{12} formula, and $g \simeq 4$ is a Gaunt factor. The value of α_{21} with χ_{21} in volts and $\chi_H = 13{\cdot}6$ volts becomes

$$\alpha_{21} = 12 \times 10^{-11}\, \chi_{21}\, T^{-1/2}. \qquad \ldots\ldots(7)$$

Combining (5) and (7), the ionization formula becomes

$$\frac{N_2}{N_1} = \frac{S_{12}}{\alpha_{21}} = 3300\, T\, \chi_{12}^{-3} \exp(-11\,600\, \chi_{12}/T). \quad \ldots\ldots(8)$$

From the fact that the ionization potentials for the coronal ions are in the neighbourhood of 400 volts we find the electron temperature is near 10^6 deg. K.

The two spectrum lines Fe XIV 5303 and Fe X 6374, whose relative intensity is used for temperature determination, are separated by four stages of ionization and require four applications of eqn. (8). The computations of Elwert (1952), Waldmeier (1952), Shklovskij (1951) and Miyamoto (1949) do not agree precisely but have been averaged to give the relation of table 6. It should be remarked,

Table 6. Ionization Temperature and Line Intensities

Intensity ratio 5303/6374	0·1	1	10
Ionization temperature (10^6 deg. K)	0·65	0·75	0·90

however, that a small localized region of abnormal temperature may not be detected by its effect on these lines since both lines would be produced by the normal regions in the same line of sight (Shklovskij 1951).

The relative populations of various ions of Fe, Ni and Ca may be measured from line intensities (Woolley and Allen 1948) and are in fair agreement with the theory.

Although eqns. (3) and (8) show the ionization to be independent of electron density, the formula should not be applied when the density is so low that no ionizing collisions take place. Let r_1 be a level such that an ion above it will not normally suffer an ionizing collision until it moves below. The ionization in higher levels will be governed by conditions just below r_1. Let l_1 be the mean movement of an ion between ionizing collisions, then at the level r_1 this should equal the scale height H. The mean free path between ionizations if the ion suffered no other collisions would be $l_j = v_i/N_1 S_{12}$, where v_i is the ion velocity and N_1 the electron density at the r_1 level. The mean free path of the ion between deflecting collisions l_d is $\mu^{1/2} l_p/Z^2$, where μ is the atomic weight, Z the nuclear charge, and l_p the proton free path given in §10. When $l_p < l_j$ a fairly simple argument leads to

$$l_1 = (l_d l_j)^{1/2}. \quad \ldots\ldots(9)$$

Equating this to H, which we may take as 4×10^{10} cm in the outer corona, one obtains $N_1 = 3 \times 10^5$ electrons cm^{-3}, and hence $r_1 \simeq 3$.

§6. Excitation

We have found that one cannot apply the Saha equation to the ionization of the corona, and for similar reasons one should not apply the Boltzmann equation to determine the distribution of atoms among the energy levels. Nevertheless a purely empirical estimate of the excitation temperature may be obtained by the usual method of plotting log (N/q), where N/q is the number of atoms per unit weight q in certain levels, against χ', the excitation potential of those levels. Then

$$\frac{5040}{T_{\text{exc}}} = -\frac{d \log (N/q)}{d\chi'} \quad \ldots\ldots(10)$$

where T_{exc} is in °K, χ' in volts, and log to base 10. Relative values of N may be found for this purpose for the upper levels of spectrum lines from the measured

intensity divided by the transition probability. The transition probabilities, excitation potentials, and statistical weights are known for all identified coronal lines (Edlén 1942, Huang 1945). Values of excitation temperatures from eqn. (10) are

	T_{exc} (°K)	Level
Huang (1945)	25000	$r=1{\cdot}1$ (?)
Woolley and Allen (1948)	13000	$r=1{\cdot}2$

It will be found that T_{exc} should decrease with height, and this probably accounts for the difference in the values quoted. As expected, T_{exc} is between the kinetic and the radiation-density temperatures.

The excitation of coronal atoms by the combined action of radiation and electronic collisions has been considered by Woolley and Allen (1948) and by Shklovskij (1950). Let the subscript 1 refer to an upper level and 0 to the ground level. In thermal equilibrium at the sun's effective temperature T, the excitation will be given by

$$\frac{N_1}{N_0}=\frac{q_1}{q_0}\exp(-\chi_1'/kT_E).$$

In dilute radiation the downward transition probability to N_0 will be almost unaffected (since stimulated transitions are negligible as compared with spontaneous transitions), but the upward transition rate will be proportional to the dilution factor $D[=\frac{1}{2}-\frac{1}{2}(1-r^{-2})^{1/2}]$, and hence in the absence of collisions the equilibrium becomes

$$\frac{N_1}{N_0}=\frac{q_1}{q_0}D\exp(-\chi_1'/kT_E). \qquad \ldots\ldots(11)$$

The actual number of downward and upward transitions per cm³ per sec in the absence of collisions would be N_1A_{10} and $N_0A_{10}(q_1/q_0)D\exp(-\chi_1'/kT_E)$ where A_{10} is the transition probability. To these must now be added the collisional de-excitation and excitation, and the transition balance now becomes

$$N_1(A_{10}+N_e\sigma_{10}v_e)=N_0[A_{10}(q_1/q_0)D\exp(-\chi_1'/kT_E)+N_e\sigma_{01}v_e]$$

where N_e is electron density, v_e is electron velocity, and σ the collision cross section. When the electron energy is much greater than the excitation energy (as in this case) we have $q_0\sigma_{01}=q_1\sigma_{10}$, where the subscript 01 implies an upward transition from level 0 to 1. The excitation is expressed by

$$\frac{N_1}{N_0}=\frac{q_1}{q_0}\frac{A_{10}D\exp(-\chi_1'/kT)+N_e\sigma_{10}v_e}{A_{10}+N_e\sigma_{10}v_e}. \qquad \ldots\ldots(12)$$

Equation (12) may be used to determine the collision cross sections σ_{10} or σ_{01}. There is no direct means of determining N_0, but one can observe the variations of N_1 as D and N_e (both known functions of r) are varied. In this way Woolley and Allen (1948) obtained the following value for the cross section of the green coronal line 5303: $\sigma_{01}=2{\cdot}3\times10^{-17}\,\text{cm}^2=0{\cdot}26\,\pi a_0^2$, where a_0 is the radius of the first Bohr orbit.

An approximate evaluation of σ_{01} may also be made by noting the change of N_1 with χ_1' for ions that emit more than one line. If we assume that σ_{10} is the same for each line and that $A_{10}\simeq100\,\text{sec}^{-1}$ in each case, the following approximate results can be obtained from the data given by Woolley and Allen (1948):

Ion	Fe XI	Fe XIII	Ni XIII	Ni XV
σ_{01} (cm²)	10^{-18}	10^{-18}	2×10^{-18}	4×10^{-18}

Calculations of the coronal line cross section by the Hebb and Menzel (1940) formula generalized by Edlén (1942) gives $\sigma_{01} = 2{\cdot}5 \times 10^{-17}$ cm², while calculations by Hill (1951) made specially for the coronal line gave $0{\cdot}11\,\pi a_0^2 = 10^{-17}$ cm². It would appear therefore that the values near 10^{-18} cm² found by relating N_1 with χ_1' are too small. Also it is found in §7 that σ_{01} should be greater than 2×10^{-17} cm² to give the expected composition of the sun's atmosphere.

§7. Chemical Composition

Spectral line photometry may be used to determine the chemical composition of the corona as of other astrophysical objects. The only elements for which identified lines are visible are Fe, Ni, Ca and A, but the method has the advantage that the results may be expressed directly in number of atoms per electron. One may assume further that there are 0·8 H atoms per electron and hence the hydrogen to metal ratio is obtainable without the need to evaluate a very large Boltzmann factor.

The intensities of coronal lines may be conveniently expressed as W, the equivalent width in terms of the electron scatter continuum. The amount of light scattered in all directions by electrons within the wavelength range W is $4\pi D I_\lambda W \sigma_e N_e$, where $4\pi D$ = the solid angle of illumination by the sun, I_λ = the mean intensity solar radiation per unit area, solid angle and wavelength range, σ_e = the electron scattering cross section, and N_e = the electron density. This equals the emitted radiation $N_1 A_{10} \boldsymbol{hc}/\lambda$, whence

$$N_1 = 4\pi D I_\lambda W \sigma_e N_e \lambda / A_{10} \boldsymbol{hc}. \qquad \ldots\ldots(13)$$

Thus $N_1(r)$ may be determined when W is measured and N_e known as a function of r. It is usually considered sufficiently accurate to use the projected value of r (as seen at an eclipse) for the true radial distance.

This procedure gives the number of atoms in the upper level. To determine the number in the ground state, and hence the number of ions, one must use the results of §6. The analysis of Woolley and Allen (1948) used an empirical excitation temperature. If instead one uses eqn. (12) it becomes necessary to adopt a value for the collision cross section. The excitation becomes

$$N_1 q_0 / N_0 q_1 \simeq D \exp(-\chi_1'/\boldsymbol{k}T) + N_e \sigma_{10} v_e / A_{10} \qquad \ldots\ldots(14)$$

as compared with the formula for an excitation temperature T:

$$N_1 q_0 / N_0 q_1 = \exp(-\chi_1'/\boldsymbol{k} T_{exc}). \qquad \ldots\ldots(15)$$

With σ_{10} equal to 2×10^{-17} cm² the values of $N_1 q_0 / N_0 q_1$ from (14) are six times smaller than from (15) for most of the coronal lines. Woolley and Allen (1948) have already shown that (15) leads to the hydrogen to metal abundance of cosmic material. Equation (14) would then lead to a six times greater proportion of metals which is unlikely in the corona. The discrepancy becomes less if one adopts a larger cross section, but cannot be removed entirely.

In order to determine the abundance of any element one must add the contributions for observed ions and, by means of (8), allow for those that emit no visible spectrum lines. The measured metal to hydrogen ratio becomes at least as high as in the lower atmosphere and gives no evidence of gravitational separation of elements. Yakovkin (1951) has estimated the abundance of ions at levels from $r = 1$ to $r = 2$ and finds very little change of composition with height.

§8. Variations with the Sunspot Cycle

It has been known for many years that the appearance of the corona varies with the progress of the sunspot cycle, but it is only in the last few years that attempts have been made to elucidate the variations of the corona's physical condition.

The appearance and main changes of the coronal plumes, streamers and arches have been described in §2. It is the distribution of streamers that dominates the character of the corona. At sunspot maximum they appear in any direction and thus give the corona a very spiked appearance which, however, looks much the same at the pole as at the equator. Soon after maximum the polar streamers disappear, but the equatorial streamers are still strong and produce an apparent ellipticity of the corona. This has often been measured by means of isophotes (Ludendorff 1928). The greatest ellipticity occurs a year or two *before* sunspot minimum (Mitchell 1929). These variations make it appear that the streamers result directly from solar activity and are superimposed on a permanent corona which would exist in the absence of any localized activity. The appearance of the permanent undisturbed corona would be a matter of some interest, but it has never been seen at eclipses. There is good evidence that the equatorial streamers are

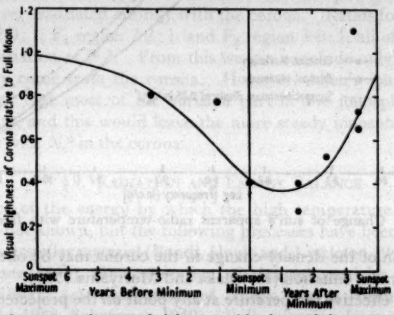


Fig. 3. Variation of corona brightness with phase of the sunspot cycle.

the visible signs of the streams of particles leaving the sun to produce recurrent geomagnetic storms (von Klüber 1952, Allen 1944). In that case they should be strongest in the equatorial direction about two years before sunspot minimum. About a year *after* minimum the equatorial streamers (judging by magnetic activity) should be the weakest and at this time the polar streamers have not started to develop. This is therefore the time when there should be the nearest approach to a completely undisturbed corona, the minimum of streamer action, and a minimum of total coronal brightness.

Photoelectric measurements of coronal brightness studied by Nikonova and Nikonova (1947) showed that the corona emitted more light at sunspot maximum than at minimum. Of the five occasions studied the faintest corona was that of 1945 July 9, one year after sunspot minimum. Saito (1950) found minimum brightness to occur one year after sunspot minimum, and a similar conclusion can be reached by plotting the visual estimates from Dyson and Woolley (1937) as in figure 3. It should be remarked that, although the 1945 July 9 corona was

about the least active ever observed, Waldmeier's observations (1947) show that the residual activity is far from negligible.

Any real change in the total coronal brightness must be interpreted as a proportional change in the total atom and electron content of the corona. The Van de Hulst (1950) analysis implies that this changes in the ratio 2:1 from maximum to minimum. It is important to investigate whether this change is in agreement with other observations.

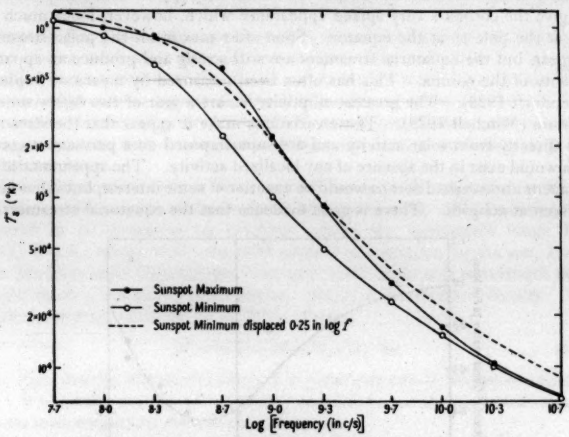


Fig. 4. Change of sun's apparent radio-temperature with frequency.

An indication of the density change in the corona may be obtained from the quiet-day solar radio emission (Hatanaka and Moriyama 1953). The emission is thermal and the effective temperature at any point on the projected disc is

$$T_{\text{eff}} = \int_0^\infty T e^{-\tau} d\tau \qquad \ldots\ldots(16)$$

where the optical depth is

$$\tau = \int_h^\infty \kappa \, dh$$

and the absorption coefficient κ is proportional to $N_e^2/T^{3/2}f^2$; h is the height and f the frequency. If throughout the sunspot cycle the corona were to change in N_e only, and if f were chosen in proportion to N_e, the calculated value of T_{eff} would remain unchanged. Thus the effect of the spot cycle would be to displace the curve of T_{eff} against log f along the log f axis without changing its shape. The apparent temperature of the sun T_a, which represents the radiation from the whole corona in terms of the sun's visible disc, would behave in the same way as T_{eff}. In fig. 4 the apparent temperature has been plotted against log f for both sunspot maximum and sunspot minimum. The maximum curve is from Allen (1951). The minimum curve is obtained by extrapolating data in the *Quarterly Bulletin of Solar Activity* to zero sunspots and combining the results with values by Piddington

(1953, unpublished). The maximum and minimum curves agree fairly well when displaced by 0·25 in log f. This implies that the electron density at maximum is 1·8 times greater than at minimum. As this agrees well with Van de Hulst's estimate we find that the radio observations can be explained as a change in electron density only.

The effect of any coronal temperature change on the radio emission would be small because the changes in thermal radiation are approximately compensated by changes in the absorption, and therefore emission, coefficient. The temperature changes measured by Waldmeier (1952) by the intensity comparison of the two coronal lines 5303 and 6374 are not large but fairly regular. At the equator the temperature minimum is nine months *after* sunspot minimum and the total variation $0{\cdot}1 \times 10^6$ deg. K. At the pole the temperature minimum is a year or two *before* sunspot minimum, the temperature maximum just before sunspot maximum, and the total variation $0{\cdot}2 \times 10^6$ deg. K. The phase of these variations is what would be expected if high temperature were associated with streamer action.

The ultra-violet emissions from the corona, whether they be line emission, free–bound or free–free transitions, are proportional to N_e^2. Consequently one would expect a variation of about $1{\cdot}8^2 : 1$, say 3 : 1, for the spot-cycle variation of any ionospheric layer associated entirely with the corona. Ratios found (Allen 1948) are E region 2·0 : 1, F_1 region 2·2 : 1, and F_2 region 3·0 : 1, all of which are fairly close to the variation of N_e^2. From this we can conclude only that ionospheric radiations *may* come from the corona. However, Allen's ionospheric analysis (1948) suggests that most of the variable part of the ionosphere comes from sunspot activity, and this would leave the more steady ionospheric variation too small to agree with N_e^2 in the corona.

§9. Radiation and Energy Balance

The source of the energy by which the high temperature of the corona is maintained is not known, but the following processes have been considered: (*a*) accretion of interstellar material (Bondi, Hoyle and Lyttleton 1947, Biermann and ten Bruggencate 1946), (*b*) acceleration of charged particles in electric fields (Shklovskij 1947), (*c*) dissipation of shock waves from solar granules (Schwarzschild 1948, Schatzman 1949), (*d*) illumination from small hot areas of the solar surface (Goldberg and Menzel 1948), (*e*) emission of high speed protons from the solar surface (Rosseland 1933, Blaha 1952), (*f*) effects of varying magnetic fields near sunspots (Giovanelli 1947, Kiepenheuer 1935), (*g*) acceleration of particles by unknown forces revealed by prominence motions, (*h*) breaking action on solar rotation due to interaction of magneto-hydrodynamic waves and the interstellar magnetic field (Leykin 1952). It is quite possible that more than one energy source is operating. Since the corona exists at all times there must be a source that is *not* associated with sunspot activity; this may be an external source or it may be associated with a continuous solar activity such as granulation. On the other hand regions near sunspots have been found hotter and denser (Waldmeier 1947, Piddington and Minnett 1951) and therefore probably represent another energy source localized near sunspots.

Since our knowledge of the source of energy is so indefinite the most promising quantitative estimates of the energy involved come from considerations of energy dissipation. Energy will be lost from the corona by conduction into the chromosphere and radiation in all directions. Probably no appreciable energy is lost

by convection because the great temperature difference between corona and chromosphere suggests that they do not intermingle.

The heat conductivity of an electron–proton plasma (Chapman and Cowling 1952, Woolley and Allen 1950) is

$$\lambda = \frac{75}{32}\left(\frac{k}{\pi m}\right)^{1/2}\frac{k^2}{e^4}T^{5/2}/A_2(2)$$
$$= 0{\cdot}5\times 10^{-6}\,T^{5/2} \qquad \ldots\ldots(17)$$

where $A_2(2) = 9{\cdot}2\log_{10}(4kT/N_e^{1/3}e^2) - 2 \simeq 50$ for the chromosphere-corona, and the other symbols have their usual meaning. The conductivity is almost independent of electron density N_e. In the presence of magnetic fields (17) will give the component of conductivity in the direction of the field. It is seen that the conductivity in the corona is enormously greater than in the chromosphere, and if the temperature of the intervening region is controlled by conducted heat there will be a very sharp delineation between the corona and chromosphere. Woolley and Allen (1950) find the energy conducted from the corona per unit sun's surface to be 2×10^4 erg cm^{-2} sec^{-1}. Alfvén (1941) and Giovanelli (1949) have obtained larger values, but it would appear that the true conduction is likely to be lower rather than higher since radio observations (Piddington 1950) suggest the corona-chromosphere temperature drop to be not as steep as obtained from the heat conductivity equation (17).

Approximate calculations have been made of the emission of various components of ultra-violet radiation. The Lyman continuum, computed on the basis of the number of recombinations in the lower corona is (Woolley and Allen 1950) 5×10^3 erg cm^{-2} sec^{-1}. The continuous emission from the main corona by free–free transitions (Woolley and Allen 1948) is 6×10^2 erg cm^{-2} sec^{-1} and the free–bound transitions 5×10^2 erg cm^{-2} sec^{-1}. Line emission from the main corona (dominated by Mg X) contributes 1×10^4 erg cm^{-2} sec^{-1} and in the lower corona (dominated by O VI) 2×10^4 erg cm^{-2} sec^{-1}. Elwert (1952) finds values of the same order for the continuous emissions, but considers (1953) that the important line emissions (about 10^3 erg cm^{-2} sec^{-1}) are in the soft X-ray region 50Å to 100Å. These various estimates are all of the right order to produce an ionospheric layer, but they cannot be computed with sufficient accuracy to enable us to recognize the sources of ionospheric radiations from solar physics.

It is evident that the quantity of high excitation energy required to maintain the corona is about 5×10^4 erg cm^{-2} sec^{-1}.

The total coronal energy content per unit solar surface as given by the kinetic energy of the protons and electrons is $E = 3kTN_0H$, where H = scale height $\simeq 10^{10}$ cm, N_0 = electron or atom density at the base of the corona $\simeq 3\times 10^8$ cm^{-3}, $T = 10^6$ deg. K and $k = 1{\cdot}38\times 10^{-16}$, E is $1{\cdot}3\times 10^9$ erg cm^{-2}. The life of the corona if the energy supply were cut off would be 6×10^4 sec or 16 hours. This shows that the main energy supply must be perpetual.

§10. Dynamical Structure of the Corona

It has been possible in earlier sections to give a fairly reliable account of the physical condition of the material of the corona. In this section we give some attention to the less understood question of the corona's structure and movement.

The plumes and arches are not difficult to comprehend. Van de Hulst (1950) has shown that the plumes may be produced in thermal equilibrium in a dipole

magnetic field and hence call for no special mechanism. The arches cannot yet be explained in precise terms, but they are built around prominences and it is reasonable to suppose that they are caused by the loss of energy or material from the corona into prominences.

The streamers form the chief visible evidence of the corona's dynamical activity. From observations they would appear to be streams of coronal material drawn from large areas of the solar surface, moving outward, narrowing into straight rays in the outer corona, and eventually being ejected into planetary space where in the earth's neighbourhood they produce geomagnetic storms. The mean sun–earth velocity is about 500 km sec^{-1} (Kiepenheuer 1947). Such a picture raises many questions and difficulties. How and where do the particles leaving the sun acquire the kinetic energy to escape from the sun and still retain 500 km sec^{-1} velocity? Does the matter move systematically outward along the streamer, in which case the outward velocity would be inversely proportional to the cross-sectional mass of the streamer? What are the actual velocities concerned? Is the outer ray caused by a jet that comes directly from the sun, or is it a consequence of a stream accelerated in the high corona? Are streamers inevitable or could the corona exist without them if the sun were very inactive? One has to assume an answer to some of these questions in order to develop a coronal theory.

There is no direct means of determining outward streamer velocity. Spectrum line shifts (Waldmeier 1947) are measured from the radiation at right angles to the stream and cannot be interpreted as a stream velocity. Studies of coronal variation during a long eclipse (Vsessviatsky, Bugoslavskaya and Deutsch 1939) would not detect motion along a smooth stream. The velocity must increase rapidly outward in order to maintain continuity of matter. If one could assume that the stream reached 500 km sec^{-1} in the outer parts of the visible corona it might be possible to deduce velocities at lower levels by eclipse photometry, but this has not yet been attempted.

The relation of prominences to streamers is open to two explanations. Since they occur within the basal area of streamers it might be thought that the primary jets causing the streamer come from the prominences (Kiepenheuer 1947). Another suggestion is that prominences are a kind of streamer backwash and are therefore located at the point where streamer motion is least.

Only rough estimates of the amount of matter being ejected in the streams may be obtained. The loss of matter by thermal evaporation is rather small to cause coronal streamers and geomagnetic storms. The mean square thermal velocity of protons at 10^6 deg. K is 160 km sec^{-1} as compared with an escape velocity of $620 r^{1/2}$ km sec^{-1}. Kiepenheuer (1951) estimates the coronal evaporation loss to be 10^8 g sec^{-1} or 8×10^8 atoms cm^{-2} sec^{-1}. By comparison the atom density in the neighbourhood of the earth on an average magnetic day might be estimated at 1 atom cm^{-3} (Ferraro 1952). Adopting a particle speed of 500 km sec^{-1}, the mean solar loss would be 2×10^{13} atom cm^{-2} sec^{-1}, about 2000 times greater than the evaporation loss. Again, Kiepenheuer's estimate is that only about 1 in 1000 atoms at $r = 3$ would be evaporated from the corona, whereas one gains the impression that most of the coronal material at this level is streaming outward. Even if one adopts a kinetic temperature of 2×10^6 deg. K at the base of the corona, evaporation loss cannot fully explain the streaming of the corona. An estimation of the coronal loss by Pickelner (1950) is 2×10^{11} atom cm^{-2} sec^{-1}, which is much closer to the geomagnetic requirement.

The dynamical structure of the corona will depend on the free path of the component protons in relation to the size of the corona as a whole. The proton cross section for collision with protons is given by the formula (Hoyle 1949)

$$\sigma_p = 8\pi (e^2/m_p v_p^2)^2 \ln(1/\theta)$$
$$\simeq 5 \times 10^{12}/v_p^4 \text{ cm}^6 \text{ sec}^{-4} \quad \ldots\ldots(18)$$

where e is the electron charge, v_p the proton velocity, m_p the proton mass, and $\theta \simeq 2 \times 10^3 (e^2/m_p v^2)$ for the corona; $\ln(1/\theta) = 14$. The free path of a proton in the corona will be approximately $1/N_e\sigma_p$. From (18) it is seen that high speed protons are much more penetrating than those at lower speeds. Kiepenheuer (1951) has shown that protons with speeds above the escape velocity can penetrate the corona from the level $r = 1{\cdot}2$ without collisions. However, the majority of protons are at the thermal velocity (160 km sec^{-1} for 10^6 deg. K). The exosphere, or level above which virtually no collisions occur, may be located by equating the free path of protons to the scale height. With $H = 4 \times 10^{10}$ cm in the high levels we have $N_e = 3 \times 10^5$ cm^{-3} and r for the exosphere about 3. An allowance for the fall of temperature with height would make the exospheric radius greater.

We are left with a picture of the corona in which most of the physical effects are associated with the high temperature. The dynamical effects are also profoundly influenced by this temperature, which however is not high enough to cause all the ejection of coronal material. A further outward force is still required to explain the coronal streamers and until such a force is understood all other dynamical problems in the corona must remain imperfectly solved.

Acknowledgments

My thanks are due to Mr. P. A. O'Brien and Dr. J. H. Piddington for sending me data in advance of publication, to the Royal Astronomical Society and the Van Nostrand Co. for use of illustrations.

It is regretted that owing to language difficulty I have not been able to make more adequate use of the numerous important Russian papers on the solar corona.

References

Alfvén, H., 1941, *Ark. Mat. Astr. Fys.*, **27**, No. 25A.

Allen, C. W., 1944, *Mon. Not. R. Astr. Soc.*, **104**, 13; 1946, *Ibid.*, **106**, 137; 1947, *Ibid.*, **107**, 426; 1948, *Terr. Mag. Atmos. Elect.*, **53**, 433; 1951, 7th *Report of Sol. Terr. Relations*, p. 63.

Bates, D. R., Fundaminsky, A., Leech, J. W., and Massey, H. S. W., 1950, *Phil. Trans. Roy. Soc.* A, **243**, 93.

Barabashev, N. P., 1939, 1936 *Eclipse Report of U.S.S.R. Acad.*, 48.

Baumbach, S., 1937, *Astr. Nach.*, **263**, 122.

Biermann, L., 1947, *Naturwissenschaften*, **34**, 87.

Biermann, L., and ten Bruggencate, P., 1946, *Nach. Akad. Wiss., Göttingen*, 19.

Blaha, M., 1952, *Bull. C.A.I. Czech.*, **3**, 29.

Bondi, H., Hoyle, F., and Lyttleton, R. A., 1947, *Mon. Not. R. Astr. Soc.*, **107**, 184.

Bugoslavskaya, E. J., 1950, *Trans. Astr. Inst. Sternberg*, **19**, 1.

Chapman, S., and Cowling, T. G., 1952, *Mathematical Theory of Non-Uniform Gases* (Cambridge: University Press), p. 179.

Dyson, F. W., and Woolley, R. v. d. R., 1937, *Eclipses of the Sun and Moon* (Oxford Clarendon Press).

Dollfus, A., 1953, *C. R. Acad. Sci., Paris*, **236**, 996.

Edlén, B., 1942, *Z. Astrophys.*, **22**, 30.

Elwert, G., 1952, *Z. Naturforsch.*, **7a**, 202, 432; 1953, *J. Atmos. Terr. Phys.*, **4**, 68.

FERRARO, V. C. A., 1952, *J. Geophys. Res.*, **57,** 15.
GARSTANG, R. H., 1952, *Astrophys. J.*, **115,** 569.
GIOVANELLI, R. G., 1947, *Mon. Not. R. Astr. Soc.*, **107,** 338; 1949, *Ibid.*, **109,** 372.
GOLDBERG, L., and MENZEL, D. H., 1948, *Centennial Symposia, Harv. Mon.*, **7,** 279.
GROTRIAN, W., 1931, *Z. Astrophys.*, **3,** 199, 220; 1934, *Ibid.*, **8,** 124; 1948, *Naturwissenschaften*, **35,** 321, 353.
HAGEN, J. P., 1951, *Astrophys. J.*, **113,** 547.
HATANAKA, T., and MORIYAMA, F., 1953, *Pub. Astr. Soc. Japan*, **4,** 145.
HEBB, M. H., and MENZEL, D. H., 1940, *Astrophys. J.*, **92,** 408.
HILL, E. R., 1951, *Aust. J. Sci. Res.*, **4,** 437.
HOYLE, F., 1949, *Some Recent Researches in Solar Physics* (Cambridge : University Press), p. 44.
HUANG, KUN, 1945, *Astrophys. J.*, **101,** 187.
KIEPENHEUER, K. O., 1935, *Z. Astrophys.*, **10,** 260; 1947, *Astrophys. J.*, **105,** 408; 1951, *Z. Naturforsch.*, **6a,** 627.
VON KLÜBER, H., 1952, *Observatory*, **72,** 207.
LAFFINEUR, M., 1952, *C. R. Acad. Sci., Paris*, **234,** 1528.
LEYKIN, G. A., 1952, *Astr. J. Moscow*, **29,** 761.
LUDENDORFF, H., 1925, *S.B. Preuss. Akad. Wiss.*, **5,** 83; 1928, *Ibid.*, **16,** 185.
LYOT, B., 1937, *Astronomie*, **51,** 203.
MACHIN, K. E., and SMITH, F. G., 1952, *Nature, Lond.*, **170,** 319.
MINNAERT, M., 1930, *Z. Astrophys.*, **1,** 209.
MITCHELL, S. A., 1929, *Handb. d'Astrophys.*, **4,** 338; 1951, *Eclipses of the Sun*. 5th Edn (New York : Columbia University Press).
MIYAMOTO, S., 1949, *Pub. Astr. Soc. Japan*, **1,** 10.
NIKONOVA, V. B., and NIKONOVA, E. K., 1947, *Pub. Crimean Astrophys. Obs.*, **1,** 83
O'BRIEN, B., STEWART, H. S., and ARONSON, C. J., 1939, *Astrophys. J.*, **89,** 26.
O'BRIEN, P. A., 1953, *Mon. Not. R. Astr. Soc.*, **113,** in the press.
PICKELNER, S. P., 1950, *Pub. Crimean Astrophys. Obs.*, **5,** 34.
PIDDINGTON, J. H., 1950, *Proc. Roy. Soc.* A, **203,** 417.
PIDDINGTON, J. H., and MINNETT, H. C., 1951, *Austral. J. Sci. Res.*. **4.** 131.
ROSSELAND, S., 1933, *Pub. Uni. Obs. Oslo*, No. 5.
SAITO, K., 1950, *Ann. Tokyo Astr. Obs.*, **3,** 1.
SCHATZMAN, E., 1949, *Ann. d'Astrophys.*, **12,** 203.
SCHWARZSCHILD, M., 1948, *Astrophys. J.*, **107,** 1.
SHAJN, G. A., 1947, *Pub. Crimean Astrophys. Obs.*, **1,** 102.
SHKLOVSKIJ, I. S., 1947, *Astr. J. Moscow*, **24,** 37; 1950, *Pub. Crimean Astrophys. Obs.*, **5,** 86, 109; 1951, *Ibid.*, **6,** 105, *The Solar Corona* (Moscow: Gosadaestevennse).
SHKLOVSKIJ, I. S., and PICKELNER, S. B., 1951, *Pub. Crimean Astrophys. Obs.*, **6,** 29.
SIEDENTOPF, H., 1950, *Ergebn. exakt. Naturw.*, **23,** 1.
SMERD, S. F., 1950, *Austral. J. Sci. Res.* A, **3,** 34.
UNSÖLD, A., 1938, *Physik der Sternatmosphären* (Berlin : Springer).
VAN DE HULST, H. C., 1950, *Bull. Astr. Insts. Netherlds.*, **11,** 135, 150.
VSESSVIATSKY, S. K., BUGOSLAVSKAYA, E. J., and DEUTSCH, A. N., 1939, 1936 *Eclipse Report of U.S.S.R. Acad.*, 25.
WALDMEIER, M., 1941, *Ergebnisse und Probleme der Sonnenforschung* (Leipzig : Becker & Erler); 1945, *Mitt. Aarganischer Naturforsch. Ges.*, **22,** 185; 1946, *Physica*, **12,** 733; 1947, *Astr. Mitt. Zürich*, Nos. 149, 151; 1948, *Ibid.*, No. 154; 1951, *Z. Astrophys.*, **29,** 29; 1952, *Ibid.*, **30,** 137; 1953+, *Die Sonnenkorona*, Vols. I, II, III (Basel : Birkhäuser).
WOOLLEY, R. v. D. R., and ALLEN, C. W., 1948, *Mon. Not. R. Astr. Soc.*, **108,** 292; 1950, *Ibid.*, **110,** 358.
WOOLLEY, R. v. D. R., and STIBBS, D. W. N., 1953, *Outer Layers of a Star* (Oxford : Clarendon Press).
YAKOVKIN, N. A., 1951, *Astr. J. Moscow*, **28,** 79.

THEORY OF ORIGIN OF COSMIC RAYS

By EDWARD TELLER

Department of Physics, University of California, Berkeley, California, U.S.A.

CONTENTS

		PAGE
§ 1.	Introduction	154
§ 2.	Sudden production or gradual acceleration of the cosmic rays	155
§ 3.	The nature of the accelerating field	156
§ 4.	The mechanism of acceleration	159
§ 5.	The energy spectrum	161
§ 6.	The acceleration process in the sun	162
§ 7.	The sources of cosmic radiation	165
§ 8.	The termination of the acceleration process	166
§ 9.	The location of the source of cosmic radiation	168
§ 10.	The gamma-ray and electron components of cosmic radiation	169
§ 11.	Secular variations in cosmic-ray intensity	171
	Acknowledgments	172
	References	172

Abstract. Primary cosmic rays consist of fast protons and heavier nuclei. The dependence of the number of cosmic rays on their energy and the relative abundance of the various nuclear species, together with the isotropic distribution of the cosmic radiation, has to be explained by any theory of cosmic rays.

It is shown that a random acceleration process, proposed by Fermi, and a physical location of the accelerating fields in radio stars, proposed by Unsöld, account satisfactorily for the main features of cosmic radiation.

§ 1. Introduction

Our atmosphere is steadily bombarded by high energy particles. Deflection of these particles in the magnetic field of the earth, collisions with nitrogen and oxygen nuclei, creation of secondary particles, radiation and decay processes give rise to a complicated series of phenomena known as cosmic rays. The primary radiation seems to be, however, quite simple. Approximately 0·3 particle arrives per second on each square centimetre from a unit solid angle. The main portion of the radiation is isotropic and independent of time. Apparent variations with direction, latitude, the hour of the day and the season of the year are mostly due to the influence of the geomagnetic field and of the atmosphere.

The greatest number of the primary particles are protons (Johnson 1938, Schein 1948). The rest are heavier nuclei (Freier, Lofgren *et al.* 1948 a, b). On the whole, the distribution of the nuclei is the same as the natural abundances of the elements (Bradt and Peters 1950). This statement holds within experimental accuracy when particles of equal e/m values are compared, and the ratio of the nuclear abundances remains unchanged while the energy per nucleon varies in a range from 10^9 ev to 50×10^9 ev (Kaplon *et al.* 1952 a, b, c). In some measurements it has been noted, however, that the nuclei of Li, Be and B are relatively more abundant than would correspond to the small natural abundance of these elements (Dainton *et al.* 1951 a, b).

At energies greater than 10^{10} ev per nucleon the energy distribution can be represented by $dE/E^{2\cdot8}$. Below 10^{10} ev fewer particles are found than would follow from this formula, and there is little evidence for energies less than 10^{9} ev. The $dE/E^{2\cdot8}$ power-law seems to hold, at least in first approximation, up to primary particles of the highest observed energies, that is up to 10^{17} ev. It is not clear, however, whether these events of extremely high energies are due to protons or to heavier nuclei.

There is no evidence that electrons or gamma-rays are present in the primary component of cosmic radiation.

A portion of the cosmic rays is time dependent and—presumably—not isotropic. Some violent solar eruptions have been followed in some instances by considerable increases in cosmic-ray intensity (Forbush 1946, Ehmert 1948, Clay 1949, Forbush, Stinchcomb and Schein 1950). The increase in intensity follows the eruption by an interval of one quarter of an hour to an hour, but only a small fraction of eruptions are followed by such cosmic-ray phenomena. The cosmic rays occurring in the wake of eruptions have a considerably lower energy than the bulk of cosmic radiation.

In addition to these marked changes, there are more gentle variations in the cosmic-ray intensity whose period coincides with that of the solar rotation (Monck and Compton 1939). These variations are again connected with the activity of sunspot regions and seem to be particularly strong from the lowest energy component of the cosmic rays (Simpson 1951).

It seems therefore clear that a relatively soft and minor component of the cosmic radiation certainly originates in the sun, and in trying to explain the source of cosmic radiation it is reasonable to look for phenomena which are similar to those occurring in active regions of the sun.

In the following we shall give some general arguments which will show that cosmic rays are probably produced by processes of random accelerations due to the interaction of individual particles with extended moving magnetic fields (Fermi 1949). We shall then discuss in some detail production of cosmic rays by the 'radio stars', which will lead to results that are in agreement with the properties of the primary component as described above (Unsöld 1949, 1951). I do not believe that this model can be proved to be correct on the basis of present information, but it certainly contains some essential elements of the true explanation of cosmic rays.

§ 2. Sudden Production or Gradual Acceleration of the Cosmic Rays

It has been suggested that the primary particles of cosmic rays are produced in some sudden or catastrophic process. Other suggestions are that cosmic radiation was created at the beginning of the universe. There are several arguments why such hypotheses seem unacceptable.

First, it would become necessary to appeal to unknown laws or unknown situations, and one should turn to such *ad hoc* hypotheses only as a last resort.

Second, the energy density of cosmic radiation is comparable with the energy density of starlight and of the matter within galaxies, including the heat energy within stars, the kinetic energy of the stars and the kinetic energy of interstellar matter. If a new type of physical process or law should generate as much energy, it is difficult to believe that this process should not manifest itself in other experiments or other astrophysical phenomena. This same

energy consideration bears upon the possibility of a uniform density of cosmic radiation throughout intergalactic spaces. If such a uniform distribution were assumed, the total energy of cosmic radiation would become greater than any other known energy within the universe with the one exception of the rest energy of all matter present. Even this last quantity would be greater than the cosmic-ray energy only by a factor of the order of one thousand. It seems much more plausible to assume that the cosmic rays are confined (Alfvén 1937) by magnetic fields to the interior of the galaxies and that these same magnetic fields shield the galaxies from exterior cosmic rays. As we shall see later, there are indications for the presence of such galactic magnetic fields.

The last and probably most powerful argument against a catastrophic or sudden production of cosmic rays is the existence of heavy primary particles. Since these have similar energies per nucleon as the protons, and since their abundance relative to that of the protons is essentially the same as the natural abundance of the corresponding elements, one should explain the proton component and the heavier component by the same hypothesis. It is extremely hard to believe that the catastrophic process in which a heavy nucleus is given a velocity of many billion volts per nucleon should not also furnish enough energy to tear the nucleus into its component parts.

It is therefore much more plausible to assume that cosmic rays are generated by a gentle acceleration which acts on all particles with similar charge-to-mass ratio (i.e. on protons and on heavier nuclei) in a similar manner.

§ 3. The Nature of the Accelerating Field

Logically, cosmic rays could be accelerated by static electric fields or by magnetic fields. The latter, of course, would have to be in motion, since in a stationary magnetic field charged particles do not gain energy. Of these two hypotheses, the first seems unlikely. It would be necessary to assume exceedingly high potential differences, and since regions of high conductivity do occur quite frequently within the galaxy, it would not be easy to maintain these electrostatic fields. On the other hand, magnetic fields are conserved for long periods, due to the effect of self-induction in a conductor of large dimensions. Ionized material in interstellar space is a conductor of this type.

In order to form a quantitative idea of the dissipation of electrostatic fields and of the permanence of magnetic fields, we consider the system of Maxwell equations:

$$\nabla \times \mathbf{E} = (1/c)\, d\mathbf{H}/dt \qquad \ldots\ldots(1)$$

$$\nabla \times \mathbf{H} = -1/c(d\mathbf{E}/dt + 4\pi\mathbf{j}) \qquad \ldots\ldots(2)$$

$$\nabla . \mathbf{E} = 4\pi\rho \qquad \ldots\ldots(3)$$

$$\nabla . \mathbf{H} = 0. \qquad \ldots\ldots(4)$$

If the measured current is in a system in which the gas is at rest and in which no 'wind' is felt, we can write

$$\mathbf{j} = \sigma\mathbf{E} \qquad \ldots\ldots(5)$$

where σ is the conductivity.

Substituting (5) into (2) we obtain

$$\nabla \times \mathbf{H} = -1/c(d\mathbf{E}/dt + 4\pi\sigma\mathbf{E}) \qquad \ldots\ldots(6)$$

from which it is evident that σ has the dimensions of a reciprocal time. The significance of this time becomes apparent if we take the divergence of both

sides of (6), neglect the derivative of σ, and eliminate $\nabla \cdot \mathbf{E}$ with the help of (3). This gives

$$d\rho/dt = -4\pi\sigma\rho \qquad \ldots\ldots(7)$$

and, therefore,

$$\rho = e^{-4\pi\sigma t}. \qquad \ldots\ldots(8)$$

Thus a charge distribution decays with a mean life of $1/4\pi\sigma$. It is to be noted that in a region in which the conductivity has a constant value, the decay period does not depend on the spatial extension of a charge distribution.

In an ionized gas the conductivity has the value

$$\sigma = (N\lambda e^2)/mv \qquad \ldots\ldots(9)$$

where N is the number of electrons per cm^3, λ the mean free path, e is the electron charge, m the electron mass and v the average velocity of the electrons. This conductivity will of course have widely varying values in the interstellar and intergalactic spaces.

Assuming that the interstellar spaces within the galaxy are filled with approximately one hydrogen atom per cubic centimetre, that this hydrogen is completely ionized and that the kinetic temperature of electrons is a few degrees absolute, one finds $v \simeq 10^6$ cm sec^{-1} and $\lambda \simeq 10^5$ cm, and therefore we get $\sigma \simeq 10^7$ sec^{-1}. Since the velocity is proportional to $T^{1/2}$ and the mean free path to T^2, the conductivity will increase as $T^{3/2}$. At higher temperatures we shall therefore expect higher conductivities. Values exceeding $\sigma = 10^{15}$ sec^{-1} might be expected.

In a completely ionized gas the product $N\lambda$ is independent of the density and thus the conductivity does not depend on the assumed number of hydrogen atoms per cubic centimetre. Hydrogen emission lines are observed to a distance of about 100 light years from hot stars of high luminosity and thus nearly complete ionization is to be expected in extensive regions within the galaxies. Even for incomplete ionization the conductivity will remain practically unchanged as long as the total scattering cross section of the ions is greater than that of the neutral atoms.

Actually, an ionization as low as one in 10^{11} will give the same product $N\lambda$ (where N is the number of *electrons*) as we obtained for complete ionization and the low temperature of a few degrees absolute. The low number of 10^{-11} electrons per proton could be easily delivered even by such rare elements as Cs or Rb. These can be ionized by radiation of 3000 Å, and such radiation is surely present throughout the galaxy. Due to ionization of some more abundant elements which are somewhat more difficult to ionize, like Na, K and Ca, it is believed that the number of electrons per cubic centimetre does not become lower anywhere within the galaxy than 10^{-5} times the number of protons per cubic centimetre. Thus we may consider that the lower bound of the conductivity in the interstellar spaces in the galaxy is near to $\sigma = 10^7$ sec^{-1}. This of course does not mean that electric fields cannot persist for longer times than 10^{-7} sec, but it does mean that extended electrostatic fields must not be expected anywhere within the galaxy.

We conclude therefore that cosmic rays cannot be generated by electrostatic fields within the galaxy. Other forms of electric fields are present and explain the acceleration of the primary cosmic rays. One possibility is to consider electric fields connected with plasma vibrations (McMillan 1950, Menzel and Salisbury 1946). The difficulty with this hypothesis is to find an adequate

mechanism to feed the required great quantities of energy into the plasma oscillations.

One can also consider the electric fields which are necessarily connected with the observed changes in magnetic fields near the sunspots and on some stars (Swann 1933). This approach seems indeed quite hopeful, but it becomes necessary to consider the motion of charged particles in the simultaneous presence of electric and magnetic fields. While electric fields will be observed in some coordinate systems, such will not be the case for the system which is moving along with the local velocity of the conducting gas. The interaction of a varying magnetic field with hydrodynamic motions in a gaseous conductor gives rise to 'magneto-hydrodynamic' phenomena which, although complicated in detail, are simple in their basic behaviour (Alfvén 1950). We shall consider first a magnetic field in a stationary conductor.

We take the curl on both sides of eqn. (6) and take the time derivative of eqn. (1). Eliminating $\nabla \times d\mathbf{E}/dt$, neglecting the derivative of σ, and using (4), one obtains the equation

$$\Delta \mathbf{H} = \frac{1}{c^2}\frac{d^2\mathbf{H}}{dt^2} + \frac{4\pi\sigma}{c^2}\frac{d\mathbf{H}}{dt} \qquad \ldots\ldots(10)$$

where the operator Δ must be considered as applying separately to each of the components of the vector $\mathbf{H}$. With the help of eqn. (10) one can understand the time dependence of $\mathbf{H}$ for three simple limiting cases which are obtained by neglecting in turn one of the three terms in the equation. In case $\Delta\mathbf{H}$ on the left-hand side is negligible, one obtains the solution $\mathbf{H} \simeq \mathbf{H}_0 \exp i4\pi\sigma t$. This is the magnetic field induced by the decaying electric field which is connected with the charge distribution given in eqn. (8). For quickly varying magnetic fields the second term on the right-hand side of (10) can be neglected compared with the first term. In this case (10) becomes the wave equation for the propagation of electromagnetic radiation.

Finally, for slowly varying fields, the first term on the right-hand side of (10) can be neglected. This is the case of greatest interest to us. It corresponds to a quasi-static magnetic field. In this case all components of the magnetic field behave according to a diffusion equation with the diffusion coefficient $(4\pi\sigma)/c^2$. According to the well-known behaviour of diffusion equations, the magnetic fields will tend to become homogeneous and will change within a period $t \simeq (4\pi\sigma L^2)/c^2$, where L is the linear dimension within which a considerable spatial variation of a magnetic field has originally been present. It should be noted that this time is proportional to σ, as indeed should be expected from the consideration that the magnetic fields are maintained by self-induction which becomes strongest in well-conducting media. If we use for σ its minimum value of 10^7 sec^{-1} and for L the distance of a light second (3×10^{10} cm), we obtain $t = 1{\cdot}2 \times 10^8$ sec $= 4$ years. If L is equal to a light day we get $t = 3 \times 10^{10}$ years, which is longer than the estimated age of the universe. Thus for distances which are quite modest on the galactic scale, magnetic fields will not be dissipated but will remain tied to a conducting gas.

This statement needs some more elaboration in case of a conducting medium which is itself in motion. Let us consider an imaginary closed loop within this conducting gas. This loop will enclose a certain magnetic flux. We think of the elements of this loop as travelling with the gas, that is as fixed in a Lagrangian coordinate system. Self-induction will maintain the magnetic flux through the

loop, no matter how the loop may be distorted in the course of its travels. This statement will not apply of course in small dimensions, but it will apply in the big dimensions considered in the previous paragraph. In the sense now specified the magnetic lines of force will be tied to the motion of conducting matter in such a way that matter can freely slide along the magnetic lines of force but is attached to the magnetic lines of force in any motion perpendicular to these lines.

Motion of magnetic lines of force through stationary matter gives rise to an electric field $E = \mathbf{v} \times \mathbf{H}/c$, where $\mathbf{v}$ is the velocity of the motion of the magnetic lines of force. The conclusion of the previous paragraph means that in the coordinate system in which the conducting matter is at rest, v is also zero and, therefore, no electric field will be felt in the system in which the conducting gas is at rest.

Because of the importance of this conclusion for magneto-hydrodynamic phenomena in the galaxy, we shall reconsider it from one more point of view. In the presence of a magnetic field, charges may flow freely along the magnetic lines of force. This means that the component of the electric field which is parallel to the magnetic field will disappear in materials of high conductivity within a very short time. If the electric and magnetic fields are perpendicular, two cases should be distinguished according to whether E or H is greater, measured in the c.g.s. system. If E is greater, a coordinate system can be found moving with velocity $c|H|/|E|$ in which the magnetic field vanishes but an electric field is present. In this coordinate system the electric field will again vanish within a very short time, and we can therefore exclude the possibility of $|E|$ being greater than $|H|$. If the magnetic field is bigger, then a coordinate system can be found moving with a velocity $c|E|/|H|$, in which a magnetic field is present but the electric field vanishes. In this coordinate system positive and negative ions will move on circles with opposite senses of rotation. Thus in this coordinate system the ions will not move relative to the lines of force of the magnetic field. Given a sufficient length of time in which ions and neutral material may come into equilibrium, a situation will be established in which no electric field is present and no 'wind' is felt in our new coordinate system. In another coordinate system both electric and magnetic fields will be present, while positive and negative ions will move in cycloids with opposite senses of rotation but equal velocities of translation.

The above discussion leads to the conclusion that changing magnetic fields will produce electric fields, but these electric fields will impart a motion $c|E|/|H|$ to the conducting matter as a whole and will not lead to strong acceleration of individual ions which could be identified with the primary particles of cosmic radiation.

§ 4. The Mechanism of Acceleration

While the electric fields produced by moving magnetic fields will not produce cosmic-ray primaries in a single process, repeated accelerations can give rise to a spectacular accumulation of energy. Fermi (1949) has shown that a great number of such collisions can give rise to the high energies characteristic of cosmic rays and that, moreover, this stepwise acceleration process accounts for the energy spectrum of the primary radiation. The theory is based on a simple but startling idea. When a charged particle is deflected by a magnetic

field which is tied to a conducting gas, there is an exchange of momentum and energy between the particle and the magnetized gas. In a single collision the charged particle may gain or lose energy, but in the long run statistical equilibrium should be approached, and this means that the particle tends to acquire an energy equal to that of the moving magnetized gas taken as a whole. The time needed to accumulate this enormous amount of energy in a single particle is too long and equilibrium is never reached. But one can understand that, on the average, acceleration of the particle continues until very high energy values are reached.

How are the particles which acquire energy to be distinguished from the other particles in the gas? The difference lies in the energy. Fast particles will be further accelerated while the slow ones continue to participate in the Maxwell distribution within the gas. We shall return to this question in connection with the injection mechanism.

Let us consider the collision of a cosmic-ray particle whose energy and momentum are E and $\mathbf{p}$, with some magnetized gas of mass M. We shall assume for the present that the gas is at rest. The rate of energy transfer depends on the average $\overline{(\delta \mathbf{p})^2}$. The only important point in our discussion is that this average is of the approximate magnitude of $\mathbf{p}^2$. Thus the cosmic ray loses the energy

$$\overline{\delta E} = -|\mathbf{p}|^2/M. \qquad \ldots\ldots(11)$$

If the conducting gas is in motion the energy exchange will contain a term proportional to $\mathbf{p} \cdot \mathbf{V}$ and also a term proportional to $\mathbf{p}^2$ where $\mathbf{V}$ is the velocity of the gas. The scalar product may be positive or negative and in first approximation its average is zero. But collisions have different probabilities according to the relative velocity of particles and gas, and, therefore, positive and negative values of $\mathbf{p} \cdot \mathbf{V}$ will be found with different frequency. The difference in the frequency of occurrence is proportional to $\mathbf{V}$. Thus the term containing the scalar product will contribute to the average change of energy a term proportional to $\mathbf{V}^2$. Thus we may replace (11) by

$$\overline{\delta E} = -\overline{\mathbf{p}^2}/M[1 - A\overline{\mathbf{V}^2}]. \qquad \ldots\ldots(12)$$

The coefficient A can be evaluated by noting that no average energy exchange should occur for thermal equilibrium. In a relativistic gas the average kinetic energy is not $3kT/2$. Instead one may set

$$\frac{c^2\mathbf{p}^2}{2E} = \frac{\int (c^2\mathbf{p}^4/2E)\, e^{-E/kT}\, d\mathbf{p}}{\int \mathbf{p}^2 e^{-E/kT}\, d\mathbf{p}} = \frac{\int \tfrac{1}{2}\mathbf{p}^3 (d\mathbf{E}/d\mathbf{p})\, e^{-E/kT}\, d\mathbf{p}}{\int \mathbf{p}^2 e^{-E/kT}\, d\mathbf{p}} = 3kT/2. \qquad \ldots\ldots(13)$$

Therefore one finds

$$\overline{\delta E} = 0 \quad \text{for} \quad \overline{M\mathbf{V}^2} = \overline{(c^2\mathbf{p}^2/E)} \qquad \ldots\ldots(14)$$

where E is the total energy of the particle. Thus we find

$$A = M\overline{(E/c^2\mathbf{p}^2)}. \qquad \ldots\ldots(15)$$

In the actual accelerating process the energy of the conducting gas mass will be very great compared with the energy of the particle. We may, therefore, neglect the first term in the bracket of (12). At high energies it is permissible to replace the average value of the ratio $E/\mathbf{p}^2$ by the ratio of the averages:

$$\overline{\left(\frac{E}{c^2\mathbf{p}^2}\right)} \simeq \frac{\overline{E}}{c^2\overline{\mathbf{p}^2}}. \qquad \ldots\ldots(16)$$

For $E-mc^2 \ll mc^2$ this is not accurate, but at these low energies no satisfactory theory of the cosmic-ray spectrum seems possible at present.

Substituting (15) and (16) into (12) we obtain

$$\overline{\delta E} = \overline{E}\mathbf{V}^2/c^2. \qquad \ldots\ldots(17)$$

We see that in this expression the velocity of the conducting gas occurs but its mass does not. It should be noted that the expression E on the right-hand side of (17) is the total energy, that is the sum of the kinetic energy and mc^2.

In this derivation several assumptions have been made which are not strictly correct. One assumption of this kind is to consider a definite mass of moving magnetized gas. Actually, the amount of matter to be attached to a moving line of force is often ambiguous. Furthermore, the energy of the cosmic-ray particle may be changed by an induction mechanism rather than by a collision mechanism. Thus the motion of the gas mass may cause a change in the local magnetic field, and this will cause an acceleration (or deceleration) of the spiralling motion of the particle. One can show, however, that all these mechanisms give rise to energy changes similar to expression (17).

In our derivation the only important thing is the tacit assumption that the process considered shall have an equal probability for all cosmic-ray particles, independent of their energy or magnetic rigidity. This assumption is fulfilled if the cosmic-ray particle is twisted around the magnetic line of force in a tight enough spiral so that within the radius of this spiral the magnetic field can be considered homogeneous. In this case deflection processes will occur similarly for all cosmic-ray particles independent of the radius of the spiral or the energy of the particle and dependent only on the pitch of the spiral and the structure of the magnetic field. Thus changes in energy will be proportional to the momentum which the particle carries ($\delta E \simeq \mathbf{p}\,.\,\mathbf{V}$). A similar proportionality of δE and $\mathbf{p}$ is found also in the mechanism of acceleration by induction. In all these cases it is possible to show that when $E-mc^2 \gg mc^2$ expression (17) is valid, while for low energies (17) seems to be a fair approximation.

The quantity $\mathbf{V}^2/c^2$ is always small compared with unity. We shall designate this quantity by α. Then it follows from (17) that in n collisions the energy will increase on the average from an original value E_0 to a value E_n given by the relation

$$E_n = E_0\, e^{\alpha n}. \qquad \ldots\ldots(18)$$

§ 5. The Energy Spectrum

We shall now assume that during the acceleration process a certain fraction of N, the number of particles, gets lost. Let the number δN disappearing per collision be

$$\delta N = -\beta N \qquad \ldots\ldots(19)$$

where β is small compared with unity and does not depend on the energy. Then the number N_n which on the average have survived n collisions will be connected with the original number N_0 by

$$N_n = N_0\, e^{-\beta n}. \qquad \ldots\ldots(20)$$

It follows from (18) and (20) (or more directly from (17) and (19)) that the number of particles with energies greater than a given energy E is proportional to the inverse β/α power of E:

$$N(E) = \text{const}/E^{\beta/\alpha}. \qquad \ldots\ldots(21)$$

Here again E is the total energy, that is the kinetic energy plus mc^2. At high energies ($E \gg mc^2$) formula (21) agrees well with the empirical results if one sets

$$\beta/\alpha = 1{\cdot}8. \qquad \ldots\ldots(22)$$

For low energies ($E < 10\,mc^2$) the number of observed particles is smaller than the number calculated from the power law. It is possible that magnetic fields near the orbit of the earth prevent some of the particles carrying smaller energies from reaching our planet. It is also possible that the cosmic-ray particles cannot escape from the neighbourhood in which they are created unless they possess a minimum amount of energy. If the latter is the case, the dependence of the number of cosmic rays on the energy at $E < 10\,mc^2$ may shed some light on the nature of the sources of cosmic radiation.

The theory of acceleration shows that the fractional energy gain per collision, that is the quantity α, is independent of the energy for $E \gg mc^2$. The fractional loss in the number of cosmic rays per collision β may well depend on the energy. The fact that a power law represents the dependence of the number of cosmic rays on the energy at $E \gg mc^2$ shows that β/α is independent of energy and that β itself must be constant. Furthermore, the same exponent seems to fit the energy distribution both of protons and of heavier nuclei. Thus the fractional loss appears to be the same for all components in the cosmic radiation.

Comparing various kinds of nuclei with equal E/mc^2 values, one finds that the relative abundance of atoms in cosmic-ray primaries is similar to the relative abundance in any representative sample of matter. One is led to assume, therefore, that any kind of nucleus will be injected with equal probability into the acceleration process. Thus both the end and the beginning of the acceleration process are governed by some occurrence which operates equally on nuclei of all kinds.

On the basis of the foregoing considerations it is easy to understand why electrons are not found in the primary cosmic radiation. Let us assume that electrons are started as frequently on the acceleration process as protons, and that the injection occurs at similar E/mc^2 values. Then the electrons will receive a smaller energy in the ratio of electron-mass/proton-mass = 0·00055. A considerable portion of the primary proton energy finds its way into showers in which numerous electrons participate. If a fraction 0·00055 of the primary energy is carried by electrons, the effects of these electrons will be hard to observe.

Actually it is to be expected that it is much more difficult to accelerate electrons than protons. For equal E/mc^2 values, and therefore for equal velocities, electrons and protons will lose similar amounts of energies to the surrounding gas particles by coulomb scattering. At the same time electrons will gain energy from the moving magnetic fields much more slowly. If we assume that initial particles are produced at the same E/mc^2 value this may well lead to effective acceleration of protons and to deceleration and effective loss of the injected electrons. This conclusion is further strengthened if one remembers that electrons may effectively lose energy to radiation as they circle around magnetic lines of force, while the same process for protons is quite ineffective.

§ 6. The Acceleration Process in the Sun

It remains to find a process which will start the acceleration of all kinds of particles with similar probabilities and to find a second process which will

terminate the acceleration with equal probabilities for all particles and for all energies. One may look for a model of these processes to the solar activities which are sometimes accompanied by bursts of radiation similar to cosmic rays.

There is evidence for magnetic fields of a few thousand gauss in the neighbourhood of sunspots, which presumably extend to considerable heights above the surface of the sun. At these greater heights, that is in the region of the corona, one encounters quite small densities, of 10^8 atoms per cm^3 or less. In this region the energy density due to the magnetic field is apt to be considerably greater than the energy density present as hydrodynamic energy and thermal energy per unit volume carried by the material. Similarly, the Maxwell tensions due to the magnetic field are greater than the material pressures which would be available to drive a purely hydrodynamic motion. Numerically, for a field of 1000 gauss the magnetic energy density and the Maxwell stress will be $H^2/8\pi = 4 \times 10^4$ erg cm^{-3}. Assuming a high velocity of 10^8 cm sec^{-1} and a density of 10^8 protons and electrons per cm^3, one obtains a kinetic energy of only 0·8 erg cm^{-3}. In order to estimate the thermal energy content and the pressure we assume a temperature $T = 10^6$ deg. K. Such high temperatures are by no means unusual in the corona. With the same particle density as above one finds for the energy density 0·04 erg cm^{-3} and for the pressure 0·027 erg $cm^{-3} = 2{\cdot}7 \times 10^{-8}$ atmosphere. Under these conditions the motion of the ionized material is almost completely determined by the magnetic lines of force. The energy, density, pressure and inertia of the material react but little on the changes of the magnetic fields. If, therefore, there is any motion or change in the magnetic fields due to some process originating in the sunspot near the surface of the sun, the tenuous material of the corona will be swept along with the moving magnetic fields. It is most probable that processes of this kind are at play in the production and development of solar prominences. It is easy to see that the gas in the corona will attain very high velocities. Actually, collisions of such rapidly moving gas masses are likely to account for the high temperatures in the corona which have been mentioned above.

Particularly rapidly moving magnetic fields may accelerate matter in the corona to velocities sufficiently high, so that further acceleration by random collisions can then transform the particles of this gas mass into cosmic rays. As an explicit example let us consider protons of a velocity 10^8 cm sec^{-1} or 5 kev kinetic energy. Actual velocities of this order do occur in disturbed regions of the sun. In a medium of 10^8 ionized hydrogen atoms per cm^3, protons of this velocity would lose their energy in approximately 10^9 cm, that is a distance comparable with the dimensions of the earth. The energy E for such protons is 10^9 ev. Assuming for the random velocity of moving magnetic fields 10^8 cm sec^{-1}, we find for the fractional energy gain $\alpha = 10^{-5}$. This gives $\delta E = \alpha E = 10$ kev. If we now imagine that some ionized gas is accelerated to 10^8 cm sec^{-1} by the magnetic field, that a fraction of this gas finds itself at the end of the acceleration process in a region of very low magnetic intensity and that this gas now impinges upon another gas mass in the corona, we shall expect that the protons which had originally belonged to the first gas mass will be accelerated further if they make a new collision with moving magnetic fields in a distance less than about 10^9 cm. It should be noted that these relatively slow protons have little magnetic rigidity and will therefore be deflected by small magnetic fields. It is needed therefore that the collisions should occur with

magnetic fields of small intensity in which, however, the magnetic lines of force move with a high velocity.

It is not necessary to assume that the initial injection process into the accelerating region should occur as described above. It is possible that plasma vibrations play a role in the initial acceleration (McMillan 1950). In any case it seems possible that injection energies as low as a few kev may suffice and that from this level onward the acceleration process might resemble more and more the one described in the previous section.

The injection process for heavier nuclei could be quite similar to that for protons. The energy loss to the surrounding gas is proportional to Z^2, whereas the energy gain in a collision with a magnetic field is proportional to mc^2. Thus the ratio of these two quantities will vary as Z^2/A. This ratio is the same for protons and alpha-particles and becomes somewhat bigger for heavier nuclei. Since the energy loss is inversely proportional to the square of the velocity, while the energy gain is practically independent of the velocity, the required injection velocity will vary as $ZA^{-1/2}$. Thus for Fe the injection velocity will have to be about five times as great as for protons. Since a sufficiently high injection velocity is only one of the factors and since it is necessary to fulfil other conditions like a sufficiently small value of the magnetic field at the right time, it may not be too surprising to find that injection of heavier nuclei occurs with a similar probability to the injection of protons. In addition, it is quite possible that during relevant periods of the injection mechanism a heavy nucleus like Fe is not yet completely ionized. At a velocity of 10^8 cm sec^{-1} an Fe atom is not apt to lose many electrons. This would lower the effective Z-value and the required velocity of injection. Thus the question of the ease of injection might become a complicated one which depends on the balancing of many factors.

After a considerable number of accelerating collisions a cosmic-ray particle may find itself on a magnetic line of force which proceeds to a great distance from the sun before turn ng back toward the solar surface. It is poss ble hat the intensity increases which are observed in cosmic rays in connection with solar outbursts are due to such particles which start from the surface of the sun and spiral along a magnetic line of force which happens to be directed toward the surface of the earth. This would explain why only a few of the violent outbursts are connected with a marked increase of cosmic-ray intensity. The time that passes between the observed solar outburst and the first increase in the cosmic-ray intensity may be due to the long path that the cosmic rays take, travelling along a curved magnetic line of force rather than along the straight line taken by the light signal which arrives from the solar outburst. The diffe ence in the time of arrival of cosmic rays moving in spirals of varying pitch accounts in part for the duration of the phenomenon. Another explanation of the time of the cosmic-ray burst is that a whole bundle of magnetic lines of force is directed toward the earth's surface and that different lines of force within this bundle happen to be directed successively toward the earth.

It seems therefore possible to understand that part of the intensity which is connected with the solar activity. At the same time we have seen that in the sun processes go on which can inject positively charged nuclei into the acceleration process discussed earlier. This injection mechanism may work for nuclei of different charges in a manner which may be not too dissimilar for different kinds of nuclei.

It has been suggested (Richtmyer and Teller 1949) that all of the cosmic radiation originates in the sun. This assumption can be reconciled with the isotropy of cosmic rays only if one assumes that the cosmic-ray particles emitted by the sun are caught in a more extended magnetic field stretching out over the planetary system. Such a field might repeatedly deflect cosmic-ray particles until eventual isotropy is approximately achieved. The solar origin of the cosmic rays might seem attractive because only a very small fraction of the sun's energy would be required to maintain cosmic radiation. All other hypotheses would make it necessary to assign much greater total amounts of energy to cosmic rays because all other explanations must fill with cosmic radiation not only the neighbourhood of the earth's orbit but all the interplanetary space as well. On the other hand, the solar hypothesis runs into difficulties when trying to account for the cosmic-ray particles of highest energies. Particles with more than 10^{14} ev per nucleon have been observed. While such particles are accelerated on the solar surface, fields of 1000 gauss extending over at least 3×10^8 cm will be needed to confine them. In the later phase, when these particles circulate in the interplanetary space, fields of 10^{-5} gauss extending over at least 3×10^{16} cm or 10^6 light seconds will be needed. This would extend the orbits of such cosmic-ray particles to the outermost limits of the planetary system.

A further argument against the solar origin of the cosmic rays is the fact that the additional radiation which accompanies solar outbursts is considerably less energetic than the steady component of cosmic radiation. It seems that in the sun the ratio β/α which determines the energy distribution of cosmic rays is greater than in the main and steady component of the cosmic radiation.

§ 7. The Sources of Cosmic Radiation

It is appropriate to look farther afield for the origin of cosmic rays. The attractive proposal that cosmic radiation should be generated over all the galactic system (Fermi 1949) suffers from the drawback that it cannot explain the presence of the heavy nuclei. According to this proposal, the acceleration of cosmic rays terminates by collision of a cosmic-ray particle with a proton in interstellar space. Collisions with these protons might well account for the behaviour of the proton component of cosmic rays. The heavier nuclei found in the cosmic radiation would have, however, a bigger collision cross section. They would be accelerated in a similar way to the protons but would be withdrawn from the acceleration process with a greater probability. This withdrawal would happen so rapidly that one could hardly account for their presence at all. Thus the similar general behaviour of the proton component and the heavy component cannot be explained on this basis.

More recently it has been proposed (Biermann 1953, Fermi 1954, Morrison, Olbert and Rossi 1954) that part or all of the acceleration process does take place in the galaxy, but that the termination of the acceleration process is escape from the galaxy, that is a process similar to the one discussed in connection with the solar acceleration. This assumption is not in disagreement with any known facts. It seems, however, that the galactic magnetic fields have a rather regular structure and are not in very strong motion. A more violently agitated magneto-hydrodynamic field would be a more plausible source of cosmic radiation.

The solar disturbances which coincide with the intensity increases in cosmic rays are also associated with the emission of radar waves of a few metres wavelength. The sun, however, is not the only celestial body which emits such radar waves. Definite positions in the sky called 'radio stars' emit radiation similar to that which is emitted by active spots on the sun. In the time average the radar waves obtained from these distant sources are greater in intensity than the waves arriving from the sun. It is reasonable to suggest (Unsöld 1949, 1951) that the radio stars in addition to emitting radar waves should also produce energetic particles similar to those which the disturbed sun emits. The greater average radar intensity emitted by these 'stars' might give reason to expect a greater flux of cosmic-ray particles arriving from them. Thus it seems likely that the main component of cosmic radiation is due to phenomena similar to solar magnetic activities but actually taking place in distant astronomical objects.

§ 8. The Termination of the Acceleration Process

Let us consider then a radio star and let us assume that it is a region in space in which magnetic fields and the hydrodynamic motion of ionized gas masses are in mutual and sometimes violent interaction. We may continue to think of an active sunspot and the adjacent regions of the corona as a small-scale model of such a radio star.

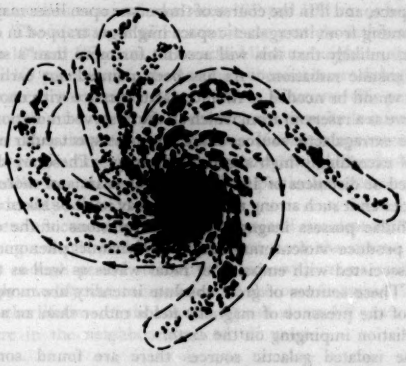
It is known that occasionally a mass of gas is expelled from the neighbourhood of a sunspot. Such a piece of plasma moves away from the sun at a high velocity, and in doing so it carries the magnetic fields with it that are frozen into that particular chunk of matter. It is plausible that similar ejections will occur from the surface of a radio star. The expelled gas mass will carry along not only its magnetic field but also the cosmic-ray particles entrapped in this field. In moving away, the gas will be dissipated, the magnetic energy will turn into kinetic energy of matter and the cosmic-ray particles will eventually escape. They are then ready to start on the journey through interstellar space.

Here we have now a mechanism by which the acceleration process can terminate with the same probability for all kinds of nuclei and for all energies. The validity of a power law with a fixed exponent which is similar for all nuclei has indicated just such a mechanism.

A slightly different mechanism accomplishing the same result may be constructed by assuming that some fraction of the magnetic lines of force originating in the radio star proceeds out into interstellar space and does not return to the surface of the radio star. Cosmic-ray particles along these lines of force can escape from the radio star, and the neighbourhood of these lines of force will soon be depleted of such particles. It is quite possible that at different times different lines of force will have the property of leaving the radio stars. This implies that some wide loops in the magnetic lines will get distorted and stretch out and eventually be opened toward the interstellar space, while open-ended magnetic lines of force may form new closed loops. In order to accomplish this some magnetic lines of force will have to move with high velocities in regions that are distant from the radio star. If, at these great distances, the density of matter is sufficiently small, one should indeed expect that the magnetic lines of force can move quite rapidly. This velocity may approach the velocity of light if the energy density of the magnetic field is greater than the density of rest energy in the same part of space (de Hoffmann and

Teller 1951). Thus it is possible that new open lines of force will continue to be formed and the cosmic-ray particles along these force lines can escape. In this model, as in the previous one, all the cosmic-ray particles, independent of charge or energy, will escape the accelerating system.

The motion of the cosmic rays within the galaxy will be influenced by any magnetic fields which happen to be present. Polarization of light coming to us from distant stars through clouds of dust indicates the presence of such magnetic fields (Hall and Mikesell 1949, Hall 1950, Hiltner 1949). Indeed, such magnetic fields might orient any ferromagnetic or paramagnetic dust particles, and orientation of conducting elongated bodies may account for the polarization of light (Spitzer and Tukey 1951). It is more probable, however, that the orientation is due to induced currents of rotating elongated bodies which are either metallic or have high metallic content (Davis and Greenstein 1951, Greenstein 1951). The interpretation of the experiments along these lines gives rise to a system of magnetic lines of force as illustrated in the figure. A frontal



Lines of magnetic force.

Note. The figure is an artist's conception of the magnetic lines of force in Sc NGC 5457 (M101), a Mount Wilson Observatory photograph printed in the monograph *Principles of Stellar Dynamics*, by S. Chandrasekhar 1942 (Chicago: University of Chicago Press).

view of a galaxy is shown and the magnetic lines of force are schematically indicated. The picture has been constructed with the idea in mind that magnetic lines of force may have been concentrated originally in the galactic centre and may have then been drawn out as the extending arms of the spiral dragged magnetic lines of force along with them.

It is to be expected that cosmic rays will continue to circulate for a considerable length of time within the galaxy. We shall return to this question. At present it may suffice to state that eventually the cosmic rays must be lost from the galaxy and that this loss process again must not distinguish sharply between cosmic rays of various energies and nuclear charges. This can be accomplished in ways similar to that used in the case of the radio stars discussed above. In fact,

the high energy density in cosmic radiation makes it appear quite plausible that in some regions the magnetic lines of force are not sufficient to confine the pressure of the gas and of the cosmic rays. These regions of interstellar gas may get lost from the galaxy. The alternative possibility is that some lines of force continue to stretch out into intergalactic space (Fermi 1954, Morrison, Olbert and Rossi 1954). In the course of time different magnetic lines of force may possess this property.

§ 9. The Location of the Source of Cosmic Radiation

Among the radio stars at least three kinds of objects have to be distinguished: the extragalactic radio stars, the isolated galactic sources and the centre of our own galaxy. We shall now turn to the question: which one of these is the most likely source of cosmic rays?

Spiral nebulae outside our galaxy emit radio waves. One assumes, however, that the magnetic field of our own galaxy will prevent these cosmic rays from entering our galaxy. If some of the galactic magnetic lines stretch out into intergalactic space, and if in the course of time some open lines may become loops, cosmic rays coming from intergalactic space might get trapped in our own galaxy. It is, however, unlikely that this will account for more than a small fraction of the observed cosmic radiation. As has been pointed out earlier, exceedingly great energies would be needed to fill intergalactic space with enough cosmic-ray particles to serve as a reservoir from which all the observed radiation can be drawn.

Among the extragalactic sources there are most spectacular objects emitting radar waves of exceedingly high absolute intensity. These are due to colliding galaxies, located at distances of 100 million light years and more. The fact that colliding galaxies emit such strong radar waves may itself be taken as an indication that spiral nebulae possess magnetic fields. Collisions of the conducting gas masses must produce violent magneto-hydrodynamic phenomena which can plausibly be associated with emission of radar waves as well as the emission of cosmic rays. These sources of great absolute intensity are more interesting as an indication of the presence of magnetic fields rather than an actual source of the cosmic radiation impinging on the earth.

Among the isolated galactic sources there are found some which are associated with exploded supernovae and others which seem to consist of wispy filaments in rapid relative motion. In each case there is evidence of hydrodynamic activity which may well be associated with strong magnetic fields. It does not need any stretch of imagination to make these objects sources of cosmic radiation as well. On the other hand, these sources will inject their cosmic radiation into interstellar space along specific lines of force which, as a rule, will not pass directly through our solar system. If, indeed, cosmic rays were received in a direct manner from an individual radio star of this kind, one would expect to find the radiation on the earth to have definite directional properties. Therefore the individual galactic radio stars cannot be regarded as a main source of the observed cosmic rays.

The radar waves originating from the centre of our galaxy do not have any identified origin. It may be natural to assume that these radar waves come from a collection of radio stars similar in nature to the individual radio stars within our galaxy. These objects, like stars in general, may well have a much greater density near the galactic centre, and in this way the great general intensity

issuing from the core of our Milky Way system can be readily understood. If this picture is correct, most of the cosmic rays generated within our galaxy will actually be produced near the galactic centre, and we have to consider the radio stars in the galactic centre as the main source of cosmic radiation.

Direct transmission of cosmic rays from the galactic centre to the earth along a specific line of force will again result in an isotropic radiation. This is contrary to experience. It is more likely that cosmic-ray particles circulate through the galaxy for a considerable period, perhaps returning to the galactic centre several times. This more involved path, which is also indicated by the figure, is apt to produce an isotropic distribution of cosmic rays at any one point in space.

It is not permissible, however, to assume that the cosmic-ray particles remain within the galaxy for an unlimited period. Heavy particles like Fe will suffer collisions with protons in the interstellar spaces. The collision cross section is approximately 10^{-24} cm^2. Since the density of protons in interstellar space is in the neighbourhood of 0·1 particle per cm^3 to one particle per cm^3, the mean free path of such a collision must be between 10^{25} and 10^{24} cm or between 10^6 and 10^7 light years. The lifetime of an Fe cosmic-ray particle within the galaxy will thus be between 10^6 and 10^7 years. The actual lifetime might be shorter than the estimated 10^6 or 10^7 years in case the particles encounter much higher density of matter on their recurrent return to the centre. Whatever this period is, it sets an upper limit to the duration in which cosmic rays must be lost from the galactic system by means other than nuclear collisions. If cosmic rays were to be retained in circulation for longer times, then the Fe particles would be depleted while the depletion of smaller cosmic-ray nuclei and in particular of cosmic-ray protons would be considerably less.

We must therefore conclude that cosmic rays will circulate within the galactic system for a long enough period to pass throughout the system several times but for a short enough period so that collisions with atoms within the galaxy will not strongly modify the composition of the cosmic rays. This will set a lower limit for the lifetime of the cosmic rays of about 10^8 years and an upper limit somewhere in the neighbourhood of 10^7 years.

§ 10. The Gamma-Ray and Electron Components of Cosmic Radiation

No primary gamma-ray component has been observed in cosmic rays. Our assumption that cosmic rays are generated in the centre region of our galaxy makes it possible to draw some more clear-cut conclusions about gamma-radiation. Such radiation could be produced either if electrons were accelerated or if protons or nuclei suffer meson-producing collisions. The neutral mesons to be expected in such collisions would then decay and emit gamma-radiation. We have pointed out above that the total energy given to electrons by our postulated acceleration process is quite small. Indeed, this mechanism could hardly produce much more than one-tenth of one per cent of the cosmic-ray energy, an amount which would pass unobserved.

Primary cosmic radiation impinging upon the earth's atmosphere does produce a considerable amount of secondary gamma-radiation through the mechanism of nuclear collisions and neutral meson production. One should therefore inquire whether the absence of gamma-radiation in the primary

component proves the absence or rare occurrence of similar nuclear collisions in the originating radio stars or in interstellar space.

If gamma-rays are produced, relatively deep inside a radio star, they may be degenerated and eventually eliminated by a shower-production process familiar in cosmic-ray phenomena. Protons or heavier nuclei possessing high energies will be eliminated by collisions with other nuclei more effectively. Thus we conclude that if both gamma-rays and fast nuclei are generated as cosmic-ray components inside a radio star, the former will escape the radio star more easily and will therefore reach the earth with less difficulty.

Emission of too great amounts of gamma-rays from the galactic centre would cause noticeable deviations from the isotropy of cosmic radiation. Thus it is likely that the energy carried by gamma-rays of this origin is less than one per cent of the total cosmic-ray energy. An equal amount of energy flux leaving the radio star in the form of protons and of gamma-rays may result, however, in a 100 times greater proton flux near the earth. In fact, the estimates made above show that protons might be trapped within the galaxy for approximately as long as 100 traversal times, whereas the gamma-rays will of course be lost after a single traversal. Even so, it is possible to conclude that the gamma-ray flux leaving the radio stars is not more, and probably is less, than the energy flux that leaves in the form of protons. This leads to the conclusion that the production of fast energy protons takes place only in a very superficial layer of the radio stars, so that most of the protons actually will escape rather than be stopped by nuclear collisions. This is the same result which we have reached above. If indeed protons were withdrawn from the acceleration process primarily by nuclear collisions, then a different law of loss of particles would hold for protons and for heavier nuclei, and one could not understand the similar dependence on energy of the various components in the primary cosmic rays.

One also may consider the production of charged mesons in nuclear collisions. These will be followed by the production of electrons which may escape the radio stars in a way similar to the one discussed in the case of gamma-rays. It is not easy to draw quantitative conclusions from the fact that this process does not give rise to a noticeable electron component in the primary cosmic rays. The amount of energy escaping as electron radiation from the radio star will be less than the amount escaping as gamma-rays because electrons will radiate while spiralling around the strong magnetic fields in the radio star. On the other hand, electrons which have escaped the radio star may be confined in the galaxy a longer time by the action of the weak galactic magnetic fields. In any case, the absence of an electron component is compatible with our general assumption that relatively few protons lose their energy in nuclear collisions and the greatest fraction escape from the radio star.

If meson production and subsequent gamma-ray emission should occur to a substantial amount during the travel of the cosmic rays within the galaxy, anisotropy should show up favouring radiation arriving in directions parallel to the galactic plane. This would be a less marked effect than the one mentioned in the last paragraphs. Similar considerations apply as in connection with gamma-ray production in the radio stars. Production of charged mesons would again give rise to an electron component in the primary radiation. On the basis of absence of gamma-rays and on the ground of the observed isotropy of the radiation, one may limit the number of nuclear collisions suffered by cosmic-ray protons. This limit is not rigorous, and one can only say that termination

of the proton orbit in the galaxy will not occur predominantly by nuclear collisions. Thus in this case, as in the previous one, absence of gamma-radiation serves to confirm the general process which we have assumed, but does not allow us to draw new or more stringent conclusions than those which have been stated earlier.

§ 11. Secular Variations in Cosmic-Ray Intensity

There is no evidence that the cosmic rays had in the past an intensity different from the present one. Investigation of the radioactive ^{14}C produced by cosmic rays is at least compatible with the idea that the cosmic-ray intensity has been constant for the last ten or twenty thousand years.

If, by a nova explosion or some other process, a radio star were to be formed on the same bundle of magnetic lines of force which pass through the solar system, and at a distance of perhaps 1000 light years or less from our own location, then a great increase in cosmic-ray intensity could be expected. This additional radiation would not be isotropic but, due to the complicated magnetic fields around the earth, it would probably reach, in greater or lesser intensities, our whole globe. The duration of such excessive intensity is hard to predict in the absence of knowledge of the nature and behaviour of the radio stars. We know that the supernova explosion which presumably gave rise to the radio star in the Crab nebula occurred almost 1000 years ago, and thus a lifetime of at least 1000 years might be expected for the cosmic-ray source. In this time the displacement of the source is approximately one-tenth of a light year. This distance is equal to the radius of the most energetic cosmic rays circling around the weak galactic magnetic lines of force. One may well assume that 1000 light years may be a typical duration for an intense burst of cosmic rays on the earth. The burst might last much longer if radio stars lived longer and if both the radio star and also our solar system would manage to stay on the same bundle of magnetic lines of force. This could happen if the radio star and also our planetary system would manage to drag along with them the bundle of magnetic lines of force on which they are sitting.

Evidence for a great increase of cosmic-ray intensity in the distant past might be found in the formation of some long-lived radioactive substance. No appropriate substance is known at present that could give evidence of such an occurrence. It is, furthermore, unlikely that such evidence could be found in any radioactivity unless the duration of the increased intensity was an appreciable fraction of the time past since that occurrence.

There may be, however, another and much more adventurous way in which cosmic rays of unusually high intensity could have left a lasting mark on the surface of our planet. Cosmic-ray irradiation amounts at present to approximately 0·01 times the medical tolerance for steady radioactive exposure. A one-million-fold increase of cosmic rays would have wiped out life from our planet with the exception of the deep sea. A 1000-fold increase of cosmic-ray intensity would have led to an increase of incidence of certain disease conditions and perhaps to an increase in the mutation rate. An increase in mutation rate by a factor 10 for a relatively short period of time may not in itself have a profound influence on the development of the species. It may be, however, that an essential change of progress in evolution such as the jumps between non-vertebrates and vertebrates could be accomplished only by a multiple, let us say sixfold, mutation. If the single mutation rate should be speeded up by a factor 10, the sixfold

mutation rate would be accelerated by a factor one million. Thus times of increased cosmic-ray intensity might be times of increased disease incidence but also times of great jumps in evolution. It is perhaps interesting to speculate why no observable fossils are found in the pre-Cambrian Age and why, at the beginning of the Cambrian, an incredible multiplicity of various living organisms emerged, apparently at the same time.

The centre of our galactic system, which contains many more stars than our suburban spiral arm, may be barren of life, due to an unbearably high steady exposure to cosmic radiation. On the other hand, the big and decisive steps in biological evolution on our own planet may be due to the greater dislocations produced in our genes by temporary increases in cosmic rays.

Acknowledgments

The above discussions have been based to a great extent on the excellent summary of properties and origin of cosmic rays given in the first chapters of *Kosmische Strahlung* (Heisenberg 1953). The author is greatly indebted to Professor Enrico Fermi for stimulating discussions.

References

Alfvén, H., 1937, *Z. Phys.*, **107**, 579; 1950, *Cosmical Electrodynamics* (Oxford : Clarendon Press).

Biermann, L., 1953, *Kosmische Strahlung*, Ed. W. Heisenberg (Göttingen : Springer-Verlag), Chap. I, Sect. 4, p. 47.

Bradt, H. L., and Peters, B., 1950, *Phys. Rev.*, **77**, 54, **80**, 943.

Clay, J., 1949, *Proc. K. Ned. Akad. Wet.*, **906**, 923.

Dainton, A. D., Fowler, P. H., and Kent, D. W., 1951 a, *Phil. Mag.*, **42**, 317.

Dainton, A. D., Gattiker, A. R., and Lock, W. O., 1951 b, *Phil. Mag.*, **42**, 396.

Davis, L., and Greenstein, J. L., 1951, *Astrophys. J.*, **114**, 206.

Ehmert, A., 1948, *Z. Naturforsch.*, **3a**, 264.

Fermi, E., 1949, *Phys. Rev.*, **75**, 1169; 1954, *Astrophys. J.*, in the press.

Forbush, S. E., 1946, *Phys. Rev.*, **70**, 771.

Forbush, S. E., Stinchcomb, Th. B., and Schein, B., 1950, *Phys. Rev.*, **79**, 501.

Freir, P., Lofgren, E. J., Ney, E. P., Oppenheimer, F., Bradt, H. L., and Peters, P., 1948 a, *Phys. Rev.*, **74**, 213.

Freir, P., Lofgren, E. J., Ney, E. P., and Oppenheimer, F., 1948 b, *Phys. Rev.*, **74**, 1818.

Greenstein, J. L., 1951, *Astrophysics, A Symposium*, Pt. 4, *Physics of Interstellar Matter*, Ed. J. A. Hynk (London : McGraw-Hill), p. 526.

Hall, J. S., 1949, *Science*, **109**, 166; 1950, *Pub. U.S. Naval Observatory*, **17**, Pt. I.

Hall, J. S., and Mikesell, A. H., 1949, *Astrophys. J.*, **54**, 187.

Heisenberg, W., 1953 (Ed.), *Kosmische Strahlung* (Göttingen : Springer-Verlag).

Hiltner, W. A., 1949, *Science*, **109**, 165.

de Hoffmann, F., and Teller, E., 1951, *Phys. Rev.*, **80**, 692.

Johnson, T. H., 1938, *Rev. Mod. Phys.*, **10**, 193.

Kaplon, M. F., Peters, B., Reynolds, H. L., and Ritson, D. M., 1952 a, *Phys. Rev.*, **85**, 295.

Kaplon, M. F., Peters, B., and Ritson, D. M., 1952 b, *Phys. Rev.*, **85**, 900.

Kaplon, M. F., Ritson, D. M., and Woodruff, E. P., 1952 c, *Phys. Rev.*, **85**, 933.

Menzel, D. H., and Salisbury, W. W., 1946, *Nucleonics*, **2**, 67.

McMillan, E. M., 1950, *Phys. Rev.*, **79**, 498.

Monck, A. G., and Compton, A. H., 1939, *Rev. Mod. Phys.*, **11**, 175.

Morrison, P., Olbert, S., and Rossi, B., 1954, *Phys. Rev.*, in the press.

Richtmyer, R. D., and Teller, E., 1949, *Phys. Rev.*, **75**, 1729.

Schein, M., 1948, *Phys. Rev.*, **73**, 1239.

Simpson, J. A., 1951, *Phys. Rev.*, **83**, 1175.

Spitzer, L., Jr., and Tukey, J. W., 1951, *Astrophys. J.*, **114**, 187.

Swann, W. F. G., 1933, *Phys. Rev.*, **43**, 217.

Unsöld, A., 1949, *Z. Astrophys.*, **26**, 176; 1951, *Phys. Rev.*, **82**, 857.

ATOMIC VALENCE STATES AND CHEMICAL BINDING

By W. MOFFITT
Department of Chemistry, Harvard University, Cambridge 38, Mass., U.S.A.

CONTENTS

	PAGE
§ 1. Introduction	173
§ 2. The method of atoms in molecules	174
§ 3. The properties of atoms in molecules	177
§ 4. Atomic structure	179
§ 5. The structure of the OH radical	182
§ 6. The electron pairing theory	185
§ 7. The valence state of the oxygen atom	187
§ 8. Valence state tables	190
§ 9. Excitation and bond energies	191
§ 10. Hybridization	195
§ 11. Electronegativities	197
References	199

Abstract. A very general theory of the electronic structure of molecules is presented, and it is shown how particular instances of it represent the main sources of chemical binding. A method of describing the properties of an atom in a specific molecule is outlined.

Particular attention is then paid to the electron-pairing theory of chemical binding, due to Slater and Pauling. It is demonstrated how an atom engaged in the formation of electron-pair bonds may be regarded as being in a non-stationary state. This is Van Vleck's valence state. The importance of the valence state concept is illustrated with reference to certain simple molecules. For example, the part played by the promotional energy of the oxygen atom's valence state in thermochemistry is outlined: whenever the bond additivity assumption is used, bond energies must be referred to atomic valence states rather than to their ground states. This is also important in assessing the excitation energies of certain molecular states, as is exemplified by the NH radical. Tables of valence states are drawn up which may be used in order to derive their promotional energies. The effect of orbital hybridization on the properties of atoms in molecules is illustrated. Finally, Mulliken's electronegativity scale is introduced and it is shown how this too depends on the energies of atomic valence states.

§ 1. Introduction

One of the most natural questions to ask of any theory of molecular structure is for the description it offers of the conditions of an atom in a particular molecular environment. How does an atom change from its isolated stationary state when it becomes the constituent of a molecule? This is not easily answered. Probably the greatest difficulty is to decide just what may be expected of such a description.

If it is agreed to adhere as closely as possible to the operational view, then there is only one satisfactory course: since nuclear motions, such as vibrations and rotations, are much slower than electronic motions, a particular atom may always

be associated with a particular nucleus in a molecule. More it is impossible to say. The conditions of an atom in a particular molecular environment are then completely specified when the electric and magnetic potentials due to all other molecular charges, together with their successive derivatives, are given in the immediate vicinity of the atomic nucleus in question. If this nucleus has an electrical asymmetry, or merely a non-vanishing spin, then the specification may often be varified experimentally. Such a description of an atom, or rather of an atomic nucleus, in a molecule is very attractive for many purposes and will indeed satisfy most physicists. But to a chemist, who may be interested in the ionization potentials of a series of ketones or the successive removal of hydrogen atoms from methane, it will not be very enlightening. His expectations are less restrictive and rather more difficult to satisfy.

It is in this connection that approximate solutions of molecular problems are particularly useful. For, quite apart from their quantitative utility, they may furnish sets of descriptive terms by means of which, for example, the properties of atoms in molecules may be defined. Each approximation will offer its own description, and the validity of each description is determined by the accuracy of the approximation with which it is associated. Care must be taken that the properties which are defined in this way should be uniquely specified and that, as an approximation is refined, their definitions should be re-examined.

Many approximate theories of molecular structure have been advanced since the introduction of quantum mechanics. The greater part of these may be related to a method which has been called that of 'atoms in molecules'. Although many authors have emphasized aspects of their respective theories which are foreign to the latter method, their approximations are most easily systematized with reference to it. Moreover, this relationship is apparently best suited to the description of chemically bonded atoms, which is the subject of the present review.

Accordingly, the discussion of atomic valence states is prefaced by two sections describing rather generally the method which gives rise to the special concepts that are to concern the later sections. This introduction has also been inserted so that the remainder of the article may be devoted to the detailed description of particularly simple illustrative examples. It is hoped that such a course will enable those who do not wish to follow the more general arguments to omit these first sections and to begin with §4. On the other hand, it will be obvious to those who are prepared to follow through the prefatory theory how the specific results in the later sections may be generalized.

§2. The Method of Atoms in Molecules

Consider two atoms x and y, and let these be brought together to form a molecule xy. The central problem of molecular structure may be described as following the state of the composite system from that of the infinitely separated atoms to those internuclear distances which characterize the stable molecule xy. Divorcing this from the problem of atomic structure, it may be assumed that accurate atomic eigenfunctions Φ_x^{ξ} are available for the description of all the states ξ of atom x, and similarly for y. Accordingly it is possible to construct mathematically well-defined functions

$$\Psi(\xi:\eta) = \{(n_x+n_y)!\, n_x!\, n_y!\}^{-1/2} \mathscr{A}\, \Phi_x^{\xi}(1, \ldots, n_x)\, \Phi_y^{\eta}(n_x+1, \ldots, n_x+n_y) \quad \ldots\ldots(2.1)$$

of the electronic spin and spatial coordinates, and of the internuclear distance r. These so-called composite functions are formed by antisymmetrizing the product of some n_x-electron eigenfunction of atom x and some n_y-electron eigenfunction of atom y. The antisymmetrizer $\mathscr{A} = \Sigma_i v_i \mathscr{P}_i$, where $\mathscr{P}_i$ is a typical permutation operator of the $(n_x + n_y)$ th degree symmetric group and $v_i = \pm 1$ according as $\mathscr{P}_i$ is even or odd respectively, has been multiplied by a suitable factor so that the composite functions are normalized asymptotically (i.e. for large values of r). At infinite internuclear distances the composite function $\Psi(\xi:\eta)$ describes accurately a system of two isolated atoms in states ξ and η respectively. Regarding the mutual approach of these atoms as a perturbation, it is clear that the complete set of composite functions may be chosen as a representation of the perturbed system, for which the internuclear distance is finite. Hence, for any given value of r it is possible to describe a particular state K of the molecule xy by the eigenfunction

$$\Omega_K = \sum_{\xi}^{(\infty)} \sum_{\eta}^{(\infty)} \Psi(\xi:\eta)\, \Gamma_K(\xi:\eta), \qquad \ldots\ldots (2.2)$$

where the expansion coefficients Γ_K are, of course, functions of r.

This formal description of the molecular problem, in which the approach of two or more atoms is regarded as a perturbation of the isolated states of these atoms, may be said to define the method of atoms in molecules. If does not, however, lead practicably to the immediate solution of any problem. This is due to two reasons. Firstly, if the set of composite functions is to be complete, the summation or integration in (2.2) must be taken over the entire spectrum, discrete and continuous, of the system of two isolated atoms. And, secondly, an accurate set of atomic functions Φ is not in general at hand. These difficulties may be partially overcome by the adoption of two artifices.

Since a practicable series method requires rapid convergence, the representation is augmented by the introduction of ionic terms. Thus the set of composite functions $\Psi(\xi:\eta)$ is supplemented by the two sets $\Psi(\xi^-:\eta^+)$, $\Psi(\xi^+:\eta^-)$ describing ionic dissociation products—or, if necessary, additional more highly ionic terms. Together, these functions are more than complete, and it is, at least formally, possible to represent the members of one set as linear combinations of the members of any other. However, if it is decided to use only a finite representation,

$$\Omega_K = \sum_{\xi}\sum_{\eta} \Psi(\xi:\eta)\, \Gamma_K(\xi:\eta) + \sum_{\xi^-}\sum_{\eta^+} \Psi(\xi^-:\eta^+)\, \Gamma_K(\xi^-:\eta^+) + \sum_{\xi^+}\sum_{\eta^-} \Psi(\xi^+:\eta^-)\, \Gamma_K(\xi^+:\eta^-), \qquad \ldots\ldots (2.3)$$

there is good evidence to show that a relatively small number of terms will ensure a rapid convergence of (2.3) to the true molecular eigenfunction in those cases where (2.2) converges only slowly.

The second device concerns the use of approximate atomic and ionic eigenfunctions $\bar{\Phi}$ in place of their accurate counterparts Φ, which are generally unknown. It has been demonstrated that the direct substitution of, for example, orbital approximations to the Φ in the definition of the composite functions leads to a molecular representation which converges much more slowly than a suitably chosen series (2.3) (cf. Moffitt 1951 a). This is due to the fact that the concomitant errors in the theoretical assessment of those purely atomic term values, associated with the infinitely separated atoms, are of the same order of magnitude as the dissociation energies which it is designed to estimate. On the other hand, these atomic term values are accessible by other means, such as the empirical study of

atomic spectra and calculations whose complexity would prohibit their extension to molecular problems but which are feasible for cases of spherical symmetry. It may therefore be assumed that these purely atomic energy levels are known. Moreover, it may be supposed that approximate atomic eigenfunctions Φ are available which reproduce the appropriate electron densities with a fidelity sufficient for the computation of interatomic forces. On this basis it has been demonstrated how approximate eigenfunctions of the form (2.3) may be calculated, together with the associated energy levels (Moffitt 1951 b).

Accepting the series expansion (2.3) as the basis of the molecular theories to be discussed, the various limiting forms which this representation may assume will furnish descriptions of the various sources of chemical binding which may arise. (In what follows it will be assumed that spin–orbit and spin–spin interactions are negligible, and the discussion will be particularly directed at the lighter elements of the periodic table.)

(i) First suppose atoms x and y to have ^{1}S ground states and include no excited or ionic terms of either atom. The series reduces to a single term, namely $\Psi(^1\mathrm{S}:^1\mathrm{S})$. As a first approximation this may be taken to represent the interaction of two rare gas atoms. The mutual approach of the atoms results in an exponentially increasing repulsion, and no primary source of binding is found.

(ii) Next, suppose atoms x and y to have multiplicities $(m+1)$, $(n+1)$ respectively, but no orbital angular momentum, again considering no excited or ionic states. The series (2.3) now contains $(m+1)(n+1)$ terms and gives rise to molecular states of multiplicities $(m+n+1)$, $(m+n-1), \ldots, |m-n|+1$ respectively. The initial spin degeneracy is partially removed by the interatomic forces in such a way as to stabilize the states of minimum multiplicity. The simplest example is the familiar hydrogen molecule where the doublet ground states of the constituent atoms give rise to an attractive singlet and to a repulsive triplet state of the molecule. It was the treatment of this molecule by Heitler and London in 1927 that launched the quantum theory of the chemical bond. The more general case envisaged above formed the basis of the Heitler–Rumer theory of spin valence (1931). Essentially it showed that if the isolated atoms in their ground states were degenerate, or contained arbitrary 'phases', then these 'phases' could be more or less favourably oriented with respect to each other under the influence of the perturbation which brought them together. When this degeneracy is orbital in origin, a similar splitting arises which it is again convenient to regard as the result of an 'orientation of phases'.

(iii) Now consider the case where the electron affinity of atom x is high and the ionization potential of y is low. Then the asymptotic separation of certain levels $\Psi(\xi^-:\eta^+)$ and $\Psi(\xi:\eta)$ is small and we may suppose that the contributions of the former are large for small r, since these will be associated with strong primary Coulomb forces of binding. This will, of course, represent the case of ionic binding.

(iv) Returning again to the rare gases, it is possible to get a better description of their interaction by the inclusion of composite functions representing their excited (Rydberg) states. This leads to the well-known dispersion forces analysed by London in 1930, which explain the source of the van der Waals attractions between such atoms.

(v) Finally, consider atoms whose outer shells are incompletely filled. Then, in addition to the composite functions representing the two atoms in their ground

states—since these are generally degenerate, there will be a number of such composite functions—the representation may be enlarged by the addition of terms referring to sub-Rydberg excitations, in which the electrons associated with the individual atoms suffer no change in total quantum numbers. It is this case that has led to what are perhaps the more fruitful concepts of chemical valence. The treatment of this approximation—for any restricted representation defines some approximation—is due particularly to Slater (1931) and to Nordheim-Pöschl (1936). It includes, as special instances, the Heitler–Rumer theory and the electron-pairing theory which is to be described below (Slater 1931, Pauling 1931).

In general, of course, it will be impossible accurately to describe a particular molecular state by means of any one of the limiting cases listed above. However, it is believed that a combination of the last-mentioned sub-Rydberg representation, supplemented by ionic terms, is sufficient to explain the main features of chemical binding among the lighter elements, and to furnish adequate estimates of molecular dissociation and excitation energies.

It should be mentioned that most presentations of the chemical theories adumbrated above have, for reasons either tacit or deliberate, identified the approximate atomic eigenfunctions, invariably of orbital type, with the accurate functions Φ. Thus, in place of the present composite functions, approximations to these, namely the

$$\widetilde{\Psi}(\xi:\eta)=\{(n_x+n_y)!\,n_x!\,n_y!\}^{-1/2}\mathscr{A}\widetilde{\Phi}_x{}^{\xi}(1,\ldots,n_x)\widetilde{\Phi}_y{}^{\eta}(n_x+1,\ldots,n_x+n_y), \qquad \ldots\ldots(2.4)$$

have been made the basis of their representations—or, more precisely, the orbital representations employed have been equivalent to the use of this basis. Since each molecular approximation corresponds to the selection of some particular set of (approximate or accurate) composite functions, and there is a one–one correspondence between the members of Ψ and of $\widetilde{\Psi}$, each is easily systematized under the above headings. As has been mentioned above, the distinction between the use of accurate and the use of approximate atomic functions is non-trivial for purposes of calculation, if not for those of description. In order to estimate effects quantitatively, it is certainly better to use the Ψ as basis. But, as will appear, just as the specification of atomic energy levels by S and L is fruitfully augmented by a specification of the associated orbital picture of the origin of these levels, so a richer description of some molecular state Ω_K is obtained by use of the correspondence $\Psi\sim\widetilde{\Psi}$.

§ 3. The Properties of Atoms in Molecules

Now it may be asked what description of the properties of a particular atom x in a molecule xy, at an internuclear distance r, this method offers. An answer may be given in the following manner.

First it is necessary to specify the range of terms which have been included in the series (2.3), thereby defining a particular approximation. Next, Ω_K is normalized in such a way that

$$\sum_{\xi}\sum_{\eta}|\Gamma_K(\xi:\eta)|^2+\sum_{\xi^-}\sum_{\eta^+}|\Gamma_K(\xi^-:\eta^+)|^2+\sum_{\xi^+}\sum_{\eta^-}|\Gamma_K(\xi^+:\eta^-)|^2=1; \qquad \ldots\ldots(3.1)$$

this is purely a matter of convenience. It may then be imagined that the molecule xy, in the state K, at an internuclear distance r, is subjected to a non-adiabatic

dissociation process, throughout which the expansion coefficients Γ_K retain for all values of r, however large, those values which they have been found to possess at r. In this way the infinitely separated atoms are prepared in a certain complex, non-stationary state which may be said to contain, in potency, all the properties of the composite system at the internuclear distance r. This dissociation will not in general correspond to any feasible physical process, but is very suggestive for the description of the molecule in terms of the properties of its constituent atoms. It will be noticed that the orthonormality of the composite functions at infinite internuclear separation leads to a representation of this non-stationary state which is normalized to unity.

The molecular function having the same analytical form when representing the isolated atoms as when it represented the results of their interaction, a statistically complete description of the properties of this non-stationary state is available. Let the charge of atom x be assessed. Then the probability that it be neutral as the result of this measurement is simply $\Sigma_\xi \Sigma_\eta |\Gamma_K(\xi:\eta)|^2$. If, indeed, the atom is found to be neutral, it is possible to associate just n_x electrons with it which may, without loss of generality, be labelled 1, 2, ..., n_x—electron exchange between the infinitely separated atoms being infinitely improbable. Immediately after this measurement, therefore, the state of the composite system is represented by the normalized function

$$\{\sum_\xi \sum_\eta |\Gamma_K(\xi:\eta)|^2\}^{-1/2} \sum_\xi \sum_\eta \Phi_x^\xi(1, \ldots, n_x) \Phi_y^\eta(n_x+1, \ldots, n_x+n_y) \Gamma_K(\xi:\eta). \quad \ldots\ldots(3.2)$$

If, further, a measurement is made of the energy of atom x, the probability of obtaining a particular result w_x^ξ, say, is given by

$$\sum_\eta^\dagger \sum |\Gamma_K(\xi:\eta)|^2 \Big/ \sum_\xi \sum_\eta |\Gamma_K(\xi:\eta)|^2, \quad \ldots\ldots(3.3)$$

where the summation $\Sigma_\xi^\dagger$ is to be taken over only those atomic states ξ (components of the degenerate level) with energy w_x^ξ. After such an observation, the system is left in the state represented by the normalized function

$$\{\sum_\xi^\dagger \sum_\eta |\Gamma_K(\xi:\eta)|^2\}^{-1/2} \sum_\xi^\dagger \sum_\eta \Phi_x^\xi(1, \ldots, n_x) \Phi_y^\eta(n_x+1, \ldots, n_x+n_y) \Gamma_K(\xi:\eta). \quad \ldots\ldots(3.4)$$

The spectral decomposition at x may then be completed, for example by the application of a magnetic field, when it is found that the probability of finding it in a particular state ξ is given by

$$\sum_\eta |\Gamma_K(\xi:\eta)|^2 \Big/ \sum_\xi^\dagger \sum_\eta |\Gamma_K(\xi:\eta)|^2. \quad \ldots\ldots(3.5)$$

The overall probability of finding atom x in a neutral state of energy w_x^ξ in the particular condition ξ is therefore given by the product of these three probabilities, namely by $\Sigma_\eta |\Gamma_K(\xi:\eta)|^2$. Similar predictions may be made for the results of measurements conducted at y, and for those which are taken first at x and subsequently at y.

It will be noticed that, so far as observations at x are concerned, the system behaves as if this were in the complex, non-stationary state

$$\sum_\xi \Phi_x^\xi G_x^\xi + \sum_{\xi-} \Phi_x^{\xi-} G_x^{\xi-} + \sum_{\xi+} \Phi_x^{\xi+} G_x^{\xi+}, \quad \ldots\ldots(3.6)$$

where, for example, $|G_x^\xi|^2 = \Sigma_\eta |\Gamma_K(\xi:\eta)|^2$; and similarly for atom y. It is therefore said that the fractional negative charge associated with atom x in the

molecular state K at internuclear separation r is given by

$$(\sum_{\xi^-}|G_x^{\xi^-}|^2-\sum_{\xi^+}|G_x^{\xi^+}|^2).$$

More generally, it may be said that in this molecular environment the atom x is in the non-stationary condition (3.6). (It will be noticed that the phase factors ascribed to the G_x's are undetermined by this decomposition. Moreover, the loss of the arguments of the Γ_K shows that the specification of the separate conditions of atoms x and y in a particular molecular environment, by means of equations (3.6), does not uniquely determine the molecular state Ω_K of (2.3)).

It will be found that particular approximations will lead to certain simple relations between the G_x's and that these may be represented pictorially and quite simply by use of the orbital theory of atomic structure. In particular, the useful concept of atomic valence states will emerge in just such a way. To this end the function (3.6) is described by its approximate form

$$\sum_{\xi}\tilde{\Phi}_x^{\xi}G_x^{\xi}+\sum_{\xi^-}\tilde{\Phi}_x^{\xi^-}G_x^{\xi^-}+\sum_{\xi^+}\tilde{\Phi}_x^{\xi^+}G_x^{\xi^+}, \qquad \ldots\ldots(3.7)$$

which may be used to embroider the prescription (3.6) in the same way that the ground ^3P state of the oxygen atom, for example, is said to arise from the electronic configuration $(1s)^2(2s)^2(2p)^4$.

However, even without such developments, a knowledge of the G_x's can be very useful. To take but two of the more obvious examples, it is often possible to correlate the fractional negative charge associated with an atom in a molecule and the chemical reactivity it exhibits; and, as Townes and Dailey (1949) have demonstrated, the greater part of the electric field gradient at nucleus x in the molecule xy arises from contributions which may be assessed on the basis of (3.6) alone; so that a knowledge of the G_x's will often enable one to calculate the electric quadrupole coupling constant of, for example, a chlorine nucleus in some chloride with fair accuracy. Conversely, a knowledge of this constant will give valuable information about the conditions under which this atom finds itself in a molecule (see, however, Moffitt 1950 b).

§4. Atomic Structure

Since the concepts of electron-pairing theory are intimately concerned with the orbital theory of atomic structure, a brief description of the latter will be given in this section using the oxygen atom as an example. The particular results obtained will be important for the illustration of specific points arising in subsequent sections. For further information on this topic the reader is referred to Condon and Shortley (1935).

When spin–orbit and spin–spin interactions are negligible, stationary atomic states (of fixed energy) may be characterized by specifying their total spin and their total orbital angular momenta, S and L respectively. A given state is then $(2S+1)(2L+1)$ fold degenerate, corresponding to the different possible components S_z, L_z of the angular momentum vectors about an arbitrarily chosen z axis; these components may assume the values $S_z=-S,\ -S+1,\ldots,\ +S$ and $L_z=-L,-L+1,\ldots,+L$. So much, and little more, it is possible to say without introducing approximations.

In order to obtain further insight into the nature and occurrence of atomic states, and in order to predict their approximate spectroscopic locations and properties, the configurational picture of atomic structure is generally adopted. The

lower states of, for example, the oxygen atom are described as arising from the configuration $(1s)^2(2s)^2(2p)^4$, in which the eight extra-nuclear electrons are assigned to specific orbitals (one-electron functions). Since the functions of 2p type are threefold degenerate, the assignment of four electrons to this shell may be made in various different ways, which are shown in the first three columns of table 1. We choose the three different 2p functions as having components of angular momentum $m_l = 0, \pm 1$ about the z axis, when they are called $2p\sigma$, $2p\pi^{\pm}$—or simply σ, $\pi^{\pm}$—respectively. Each row of the table then represents a different spatial arrangement or configuration of the electrons in the 2p levels. It is easy to identify the different atomic states which arise, using our knowledge of the degeneracies of different S, L states.

When two electrons occupy the same orbital, their total spin vanishes. However, when two orbitals contain only one electron each, then their total spin may be either 0 or 1: each orbital may be associated with an electron of either $\alpha\,(s_z = +\frac{1}{2})$ or $\beta\,(s_z = -\frac{1}{2})$ spin. And, in this way, four different components are obtained, three of which belong to a triplet $S = 1$, the remaining one being the singlet $S = 0$. These possibilities are shown in the fifth column of table 1.

Table 1. The Origin of the ^{3}P, ^{1}D, ^{1}S States from the Atomic Configuration p^4

	$2p\sigma$	$2p\pi^+$	$2p\pi^-$	Σm_l	S	States	No. of functions
I	xx	xx		+2	0	^{1}D	1
II	xx		xx	−2	0	^{1}D	1
III	x	xx	x	+1	0, 1	^{1}D, ^{3}P	4
IV	x	x	xx	−1	0, 1	^{1}D, ^{3}P	4
V	xx	x	x	0	0, 1	^{1}D, ^{3}P, ^{1}S	4
VI		xx	xx	0	0	^{1}D, ^{3}P, ^{1}S	1

For example, the four distinct functions associated with spatial configuration III are

$$\widetilde{\Phi}_{III}(S=1, S_z=+1) = |\pi^+\bar{\pi}^+\sigma\pi^-| = \widetilde{\Phi}(^3\mathrm{P}, 1, +1)$$

$$\widetilde{\Phi}_{III}(S=1, S_z=-1) = |\pi^+\bar{\pi}^+\bar{\sigma}\bar{\pi}^-| = \widetilde{\Phi}(^3\mathrm{P}, 1, -1)$$

$$\widetilde{\Phi}_{III}(S=1, S_z=0) = \frac{1}{\sqrt{2}}\{|\pi^+\bar{\pi}^+\sigma\bar{\pi}^-| + |\pi^+\bar{\pi}^+\bar{\sigma}\pi^-|\} = \widetilde{\Phi}(^3\mathrm{P}, 1, 0)$$

$$\widetilde{\Phi}_{III}(S=0, S_z=0) = \frac{1}{\sqrt{2}}\{|\pi^+\bar{\pi}^+\sigma\bar{\pi}^-| - |\pi^+\bar{\pi}^+\bar{\sigma}\pi^-|\} = \widetilde{\Phi}(^1\mathrm{D}, 1, 0) \quad \ldots\ldots(4.1)$$

(The last form of writing is explained in the next paragraph.)

We have used the abbreviation

$$|\phi_1\bar{\phi}_2\phi_3\bar{\phi}_4| = (4!)^{-1/2}\det\{\phi_1(1)\alpha(1), \phi_2(2)\beta(2), \phi_3(3)\alpha(3), \phi_4(4)\beta(4)\}; \quad \ldots\ldots(4.2)$$

bars above the orbital symbols show that these are to be taken with the spin function β, and the remaining orbitals are associated with α spin.

In the fourth column we show the value $L_z = \Sigma m_l$ of the component of the total orbital angular momentum associated with each spatial configuration. Since the highest value of L_z occurs at +2, namely in row I, for which $S = 0$, we know that

there must be a ^{1}D state. This will have four other components with $L_z = +1$, 0, -1, -2. Three of these may be identified straight away : the $L_z = -2$ component is clearly II, and those functions of III and IV with $S = 0$ are the $L_z = \pm 1$ components of the ^{1}D state. The remaining component with $L_z = 0$ is to be found in V and VI. Similarly, since there is a set of three functions with $L_z = +1$, $S = 1$, namely in III, there must be a ^{3}P state, whose three $L_z = -1$ ($S_z = 0, \pm 1$) components are to be found in IV, and whose $L_z = 0$ components arise from V. This leaves us with only one other function, namely that function with $L_z = 0 = S_z$ which arises from V and VI, and which is independent of the ^{1}D component of the same L_z, S_z: there is also a ^{1}S state. We see, therefore, that from the different possible assignments of the four 2p electrons the three levels ^{1}D, ^{1}S and ^{3}P are obtained which are respectively 5-, 1- and 9-fold degenerate. We adopt the notation $\tilde{\Phi}(^{2S+1}L,\ L_z,\ S_z)$ to distinguish the different orbital-type functions associated with these states.

The physical basis for the adoption of the orbital or configurational description of atomic states lies in the self-consistent field picture which, roughly speaking, allows each electron to move in the time-average field of the remaining electrons and of the atomic nucleus. As a qualitative guide, this approach is conspicuously successful since it enables us to enumerate and to identify the atomic states which arise, and to find their approximate spectroscopic locations. It may also be made the basis for a quantitative study of atomic states. At its most accurate, this leads to the so-called Hartree–Fock procedure, which uses the variational principle in order to determine, numerically, the radial parts of the atomic orbitals. Rather less accurate, but in some ways more useful, functions are obtained by assuming that the orbitals have a given analytical form, and determining their specific parameters by again minimizing the total electronic energy. Using either of these means, it is possible to employ equations like (4.1) to predict atomic properties quantitatively (see Hartree 1946).

It is important to realize that the set of orbital functions $\tilde{\Phi}$ so obtained does not furnish us with a set of accurate solutions Φ of the appropriate atomic eigenvalue problems. For all the states in which we shall be interested there is an obvious one–one correspondence between the members of these two sets. Thus the lowest states predicted for the oxygen atom using the above arguments based on the properties of $\tilde{\Phi}$, namely ^{1}D, ^{1}S and ^{3}P, do indeed turn out to be the most stable states. Moreover, using more detailed arguments, it is possible to forecast that the ground state will be the ^{3}P state, followed in turn by the ^{1}D and ^{1}S states—which expectation is borne out by experiment. The configurational description of atomic states, though not rigorous, is extremely fruitful and does not in general lead to any false qualitative inferences. It is less successful quantitatively, however. For example, the ^{1}S–^{3}P interval is observed to be 4·18 ev; using the full Hartree–Fock method and the best available analytical orbitals, this interval is predicted to be 5·14 ev and 6·37 ev respectively (Hartree and Swirles 1939, Moffitt 1950 a). Again, whereas both approximations lead to a $(^1\mathrm{D} - {}^3\mathrm{P})/(^1\mathrm{S} - {}^1\mathrm{D})$ ratio of about 0·67, the observed value is found to be 0·88. When dealing with more highly excited states, or with ionic species, this error is magnified quite considerably. The greater part of it is probably due to the fact that the orbital approximation does not allow the electrons to correlate the phases of their respective motions with one another, except by operation of the exclusion principle. By supposing that such correlations are permitted, the qualitative validity of the configurational picture is greatly

enhanced: the descriptive, specifically orbital concepts with which it abounds may be employed even when the term values of atomic states are prescribed with an accuracy which the Hartree–Fock method could not hope to emulate. The properties of the approximate functions Φ may be used reliably to amplify the bare statement that an atomic state, of given symmetry (S, L), occurs in some narrow region of the spectrum.

§ 5. The Structure of the OH Radical

Now let us determine what happens when we let a hydrogen atom in its ground state approach an oxygen atom in one of its states which are described as arising from $(1s)^2(2s)^2(2p)^4$. It will be convenient to suppose that the internuclear axis is the z axis, with respect to which the components of angular momenta S, L of the oxygen atom have been quantized. The resulting molecular terms are characterized by two quantum numbers S and $|\Lambda|$, of which the latter gives the magnitude of the component of the total electronic orbital angular momentum about the internuclear axis. S, which again gives the total spin angular momentum, is obtained by combining vectorially the spin vectors S_O and S_H associated with the oxygen and hydrogen atoms respectively. For the hydrogen atom we have $S_H = \frac{1}{2}$, so that for states which dissociate into a ^2S hydrogen atom and a triplet oxygen atom we have both doublet $(S = \frac{1}{2})$ and quartet $(S = \frac{3}{2})$ states. On the other hand, if the oxygen moiety is in a singlet state, it must arise from a doublet state of the OH radical. Moreover, since the hydrogen atom is to have no orbital angular momentum, it is clear that Λ may be identified with L_z of its oxygen atom dissociation product. In this way we see that the different possible molecular states which arise are those given in table 2. (For the more general treatment of such an analysis see Wigner and Witmer 1928.) So much for formal considerations. The question now arises how these states are to be described.

Table 2. The Symmetries of the Lower States of the OH Radical

Dissociation products	Molecular states
$H(^2S)+O(^1D)$	$^2\Delta$, $^2\Pi$, $^2\Sigma^+$
$H(^2S)+O(^1S)$	$^2\Sigma^+$
$H(^2S)+O(^3P)$	$^4\Pi$, $^4\Sigma^-$ $^2\Pi$, $^2\Sigma^-$

We shall use a perturbation method which presupposes that the atomic eigenfunctions of the oxygen and hydrogen atoms are known. As a first approximation we consider only those non-polar states which may be said to arise from the ground configurations of each atom. Initially, we shall deliberately eschew the orbital approach.

Our perturbation method chooses as its zeroth-order functions (in terms of which the molecular eigenfunction is to be developed) antisymmetrized products Ψ of the supposedly accurate atomic eigenfunctions Φ. Each of these basic Ψ's therefore represents a stationary state of the system of isolated atoms in the case of infinite internuclear separation. Let us see how this method may be applied in the case of the ground $^2\Pi$ state of the OH radical. This state is 2×2-fold degenerate, corresponding to $\Lambda = \pm 1$ (there being no sense of rotation about the z-axis which is particularly singled out when, as now, we are considering a free non-rotating molecule) and $S_z = \pm \frac{1}{2}$; we may therefore confine our attention to a

particular component, with $\Lambda = +1$, $S_z = +\frac{1}{2}$, say. From table 2 we see that it may arise from a ground state hydrogen and an oxygen atom in either its ^1D or in its ^3P state.

When the oxygen atom is in its ^1D state, the appropriate Ψ is easily written down. Since $\Lambda = L_z$ at the oxygen atom, and the ^1D state has vanishing spin, we must form the product of $\Phi(^1D, 1, 0)$, whose four electrons are conveniently labelled from 1 to 4, and $h(5)$, and then antisymmetrize. We use h to denote the 1s orbital of hydrogen associated with α spin, and $\bar{h}$ when the electron in this orbital has spin β. In this way we find that

$$\Psi_D(^3\Pi, 1, \tfrac{1}{2}) = \frac{1}{\sqrt{5}}\left(\mathscr{P}_1 - \sum_{i=1}^{4} \mathscr{P}_{i5}\right)\Phi_O(^1D, 1, 0)\,h(5). \quad \ldots\ldots(5.1)$$

Here $\mathscr{P}_1$ is the identical permutation, and $\mathscr{P}_{i5}$ the simple transposition which interchanges the coordinates of electrons i and 5.

Similarly, from a ^2S hydrogen and a ^3P oxygen we obtain two functions with $\Lambda = +1$, $S_z = +\frac{1}{2}$, namely

$$\Psi_1(\Pi, 1, \tfrac{1}{2}) = \frac{1}{\sqrt{5}}\left(\mathscr{P}_1 - \sum_{i=1}^{4} \mathscr{P}_{i5}\right)\Phi_O(^3P, 1, 0)\,h(5),$$

$$\Psi_2(\Pi, 1, \tfrac{1}{2}) = \frac{1}{\sqrt{5}}\left(\mathscr{P}_1 - \sum_{i=1}^{4} \mathscr{P}_{i5}\right)\Phi_O(^3P, 1, 1)\,\bar{h}(5). \quad \ldots\ldots(5.2)$$

From these two functions we must select that combination which corresponds to a doublet ($S = \frac{1}{2}$) state. The remaining function will be the $S_z = +\frac{1}{2}$ component of the quartet ($S = \frac{3}{2}$) state $^4\Pi$. Hence it is found that

$$\Psi_P(^3\Pi, 1, \tfrac{1}{2}) = \frac{1}{\sqrt{3}}(\Psi_1 - \sqrt{2}\,\Psi_2). \quad \ldots\ldots(5.3)$$

It may be remarked that the so-called composite functions Ψ so defined are normalized and orthogonal for infinite internuclear separations, but that owing to atomic overlap this condition no longer obtains when the distance between the oxygen and hydrogen nuclei is finite.

Let us now see how the perturbation method describes the mutual approach of the two atoms. When these are still very far apart, the two $^3\Pi$ states are well represented by the two functions Ψ_P, Ψ_D of (5.1), (5.3). As a first approximation, the perturbation being very small, the energies are assessed by the formula

$$E_R = \int \Psi_R^* \mathscr{H} \Psi_R \, dV \Big/ \int \Psi_R^* \Psi_R \, dV, \quad (R = P, D), \quad \ldots\ldots(5.4)$$

where $\mathscr{H}$ is the molecular Hamiltonian. These energies may be plotted as functions of the internuclear distance, and are shown diagrammatically by the dotted curves in the figure. For large distances, the energy of interaction is very small with respect to the separation of the ^1D and ^3P levels of the oxygen atom, and there is negligible interaction between these two curves: the S, L structure, which characterizes the free oxygen atom, is preserved. However, as the atoms are allowed to come into closer proximity, this representation is no longer the best which is at hand. The approximate molecular eigenfunctions are now given by linear combinations of the composite functions:

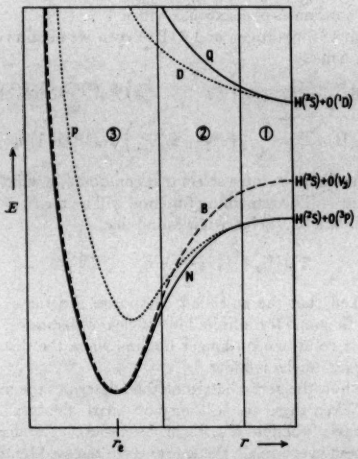
$$\Omega_K \simeq \Psi_P \Gamma_{PK} + \Psi_D \Gamma_{DK}, \quad (K = N, Q). \quad \ldots\ldots(5.5)$$

That is, the two dotted curves E_P and E_D of the figure interact to give two states N, Q, each of which contains admixtures of Ψ_P and Ψ_D. The appropriate expansion

coefficients Γ may be determined by the variational principle, which ensures that the energy

$$E_K = \int \Omega_K^* \mathcal{H} \Omega_K \, dV \Big/ \int \Omega_K^* \Omega_K \, dV \qquad \ldots\ldots(5.6)$$

assumes a stationary value. Since Ω_N, the ground state function, is no longer purely Ψ_P, it is no longer possible to say that the oxygen atom is in a ^3P state: the S, L coupling at the oxygen atom is broken down. The resulting energy curves E_N, E_Q are shown by the full lines in the figure. It is seen that the repulsion between the E_P, E_D curves leads to an increased stability for the lower state. If we now wish to describe the state of the oxygen atom in the molecule, we would have to say that it was partly in a ^3P and partly in a ^1D state (see § 7).



The validity of the electron-pairing picture of bonding in the OH radical:
(1) Domain of S, L coupling.
(2) Transition from S, L to V coupling.
(3) Domain of V coupling.

It is most desirable, for purposes of interpretation above all, to construct approximations to the composite functions Ψ and the molecular functions Ω, using the orbital atomic functions in place of their accurate counterparts Φ. These are obtained by simply substituting $\widetilde{\Phi}$'s in place of Φ's in eqns. (5.1), (5.2) and using the $\widetilde{\Psi}$'s so obtained for substitution in (5.3) and (5.5). It should be noted, however, that the Γ's of (5.5) are, in principle, determined by use of the accurate atomic and composite functions. We may then say that

$$\widetilde{\Omega}_N = \widetilde{\Psi}_P \Gamma_{PN} + \widetilde{\Psi}_D \Gamma_{DN} \qquad \ldots\ldots(5.7)$$

illustrates, in orbital terms, the function Ω_N of (5.5).

§6. The Electron Pairing Theory

A particularly enlightening formulation of the theory developed in the preceding section may now be given based not on atomic states, but on atomic orbitals and electron-pair bonds.

According to the Slater–Pauling theory (1931), a stable OH radical is formed by pairing together the spin of the electron in the 1s orbital of the hydrogen atom and the spin of the electron in that orbital of the oxygen atom which h overlaps most strongly, namely the $2p\sigma$ orbital. The formula $:\ddot{O}—H$ for the OH radical is interpreted, in this theory, by the configuration

$$(1s)^2(2s)^2(2p\pi)^2(2p\sigma \longleftrightarrow h)^2,$$

where the spins of the last two electrons are antiparallel: whenever electrons i and j occupy $2p\sigma$ and h, the associated spin function is $\{\alpha(i)\beta(j)-\beta(i)\alpha(j)\}$. Thus the function associated with this valence bond (VB) formula is

$$\widetilde{\Theta}_B = \frac{1}{\sqrt{2}}\{|\pi^+\bar{\pi}^+\pi^-\bar{\sigma}h|-|\pi^+\bar{\pi}^+\pi^-\sigma\bar{h}|\} \qquad \ldots\ldots(6.1)$$

for the $\Lambda=+1$, $S_z=+\frac{1}{2}$ component of the ground $^2\Pi$ state. By comparing (6.1) with eqns. (4.1), (5.1), (5.2), (5.3) and (5.5) it is easily seen that

$$\widetilde{\Theta}_B = -\frac{\sqrt{3}}{2}\widetilde{\Psi}_P + \tfrac{1}{2}\widetilde{\Psi}_D. \qquad \ldots\ldots(6.2)$$

Clearly, therefore, the simple VB description of the OH radical in its ground state illustrates, in the sense of (5.7), a special case of the more general perturbation formula (5.5), namely

$$\Theta_B = -\frac{\sqrt{3}}{2}\Psi_P + \tfrac{1}{2}\Psi_D. \qquad \ldots\ldots(6.3)$$

By using a qualitative criterion, namely the electron pairing picture of a chemical bond, a particular choice of the coefficients Γ_{PN}, Γ_{DN} is made. It leads to a simplification of the perturbation problem. For if we accept this idea of a chemical bond there is no need to perform an elaborate calculation to determine the coefficients Γ in (5.5).

[If, on the other hand, we were to carry through the full electron-pairing treatment (see, for example, Pauling 1933), we should also have to consider the structure which pairs together the spin of the electron in the hydrogen orbital with that of the electron in the $2p\pi^-$ oxygen orbital. In this case we have the bond function

$$\widetilde{\Theta}_C = \frac{1}{\sqrt{2}}\{|\pi^+\bar{\pi}^+\sigma\bar{\pi}^-h|-|\pi^+\bar{\pi}^+\sigma\pi^-\bar{h}|\}$$

$$= \frac{\sqrt{3}}{2}\widetilde{\Psi}_P + \tfrac{1}{2}\widetilde{\Psi}_D. \qquad \ldots\ldots(6.4)$$

The best function of this type is then a linear combination of the two bond functions, namely,

$$\widetilde{\Theta}_B\Gamma_{BN} + \widetilde{\Theta}_C\Gamma_{CN} = \frac{\sqrt{3}}{2}(\Gamma_{CN}-\Gamma_{BN})\widetilde{\Psi}_P + \tfrac{1}{2}(\Gamma_{CN}+\Gamma_{BN})\widetilde{\Psi}_D,$$

which illustrates the molecular representative

$$\Omega_N = \frac{\sqrt{3}}{2}(\Gamma_{CN}-\Gamma_{BN})\Psi_P + \tfrac{1}{2}(\Gamma_{CN}+\Gamma_{BN})\Psi_D. \qquad \ldots\ldots(6.5)$$

However, the representations (6.5) and (5.5) are entirely equivalent. The importance of the VB formalism derives from the fact that it enables us to neglect the contribution of Θ_C on the grounds that it is unreasonable to expect appreciable bonding between the $2p\pi^-$ oxygen orbital and the hydrogen 1s orbital. These do not overlap at all, whereas the $2p\sigma$ orbital overlaps h strongly.]

The validity of the simple VB formula (6.3) must now be discussed. We regard it as a qualitative criterion enabling us to write down immediately the coefficients Γ_{PN}, Γ_{DN} in the molecular function (5.5). Clearly, to begin with, Θ_B is a bad approximation when the internuclear distance is large. The asymptotic form of $|\Gamma_{PN}|$, as we have seen, is not $\sqrt{3}/2$, but 1, for the ground $^2\Pi$ state of the OH radical dissociates adiabatically into $H(^2S)+O(^3P)$, which is represented by Ψ_P. However, in the neighbourhood of the equilibrium internuclear distance r_e, calculations support the qualitative conclusion that Ω_N is represented, to good approximation, by Θ_B (cf. the calculations of Kotani 1951 as modified to suit the present approach by the methods of Moffitt 1951 b). By evaluating the energy E_B associated with the electron-pairing function Θ_B, as a function of r, we obtain a curve which is shown diagrammatically by the broken line in the figure.

It may therefore be concluded that the range of values for the OH distance may be divided up into three domains. These are indicated by (1), (2) and (3) in the figure. In (1) the interaction between the two atoms is sufficiently weak for the structure of the oxygen atom to be essentially unaltered. We may call this the domain of S, L coupling. In (3) the ground state is given, to good approximation, by the VB formula (6.3); this we call the domain of valence (V) coupling. Separating these two zones is an intermediate domain (2) which represents the gradual transition from S, L to V coupling.

The orbital representation (6.1) of the VB function Θ_B has some interesting transformation properties which we shall describe before proceeding. Suppose we were to choose a new axis z^0, inclined at an angle θ to the z axis. And, further, let us use this new axis in order to prescribe new spin functions

$$\alpha^0 = \alpha \cos \tfrac{1}{2}\theta + \beta \sin \tfrac{1}{2}\theta ,$$
$$\beta^0 = -\alpha \sin \tfrac{1}{2}\theta + \beta \cos \tfrac{1}{2}\theta .$$

Thus an electron with spin function α^0 (or β^0) would be associated with a component $+\frac{1}{2}$ (or $-\frac{1}{2}$) of its spin angular momentum about the z^0 axis. If we were to refer the quantization of the electronic spins to this z^0 axis, the electron-pairing function (6.1) would be re-defined as

$$\widetilde{\Theta}_B{}^0 = \frac{1}{\sqrt{2}} \{ |\pi_0{}^+ \bar{\pi}_0{}^+ \pi_0{}^- \bar{\sigma}_0 h_0| - |\pi_0{}^+ \bar{\pi}_0{}^+ \pi_0{}^- \sigma_0 \bar{h}_0| \},$$

where, for example, $\pi_0{}^+$ indicates that the electron in the $2p\pi^+$ orbital has spin α^0, and $\bar{\pi}_0{}^+$ that this electron has spin β^0. Expressing some of the new functions in terms of the old functions, we may easily show that

$$\widetilde{\Theta}_B{}^0 = \frac{1}{\sqrt{2}} \{ |\pi^+ \bar{\pi}^+ \pi_0{}^- \bar{\sigma} h_0| - |\pi^+ \bar{\pi}^+ \pi_0{}^- \sigma_0 \bar{h}_0| \},$$
$$= \frac{1}{\sqrt{2}} \{ |\pi_0{}^+ \bar{\pi}_0{}^+ \pi_0{}^- \bar{\sigma} h| - |\pi_0{}^+ \bar{\pi}_0{}^+ \pi_0{}^- \sigma \bar{h}| \},$$
$$= \frac{1}{\sqrt{2}} \{ |\pi^+ \bar{\pi}^+ \pi_0{}^- \bar{\sigma} h| - |\pi^+ \bar{\pi}^+ \pi_0{}^- \sigma \bar{h}| \}.$$

These identities are a simple consequence of the determinantal form of the functions concerned. We notice that, apart from the $2p\pi^-$ orbital factor, which describes the unpaired electron, the function remains unaffected by the rotation of spin axes. Whenever the spins of the electrons associated with two particular orbitals are perfectly paired, that part of the function which contains them is invariant with respect to different choices for the axes of spin quantization. We say, for example, that the quantization of the spin of the $2p\pi^-$ electron is independent of that of the $2p\sigma$ electron; or, rephrasing this statement, that the relative orientation of the spin vectors associated with the electrons in the $2p\sigma$ and $2p\pi^-$ orbitals is random. Similarly, the mutual inclination of the spin vectors associated with one of the electrons in the $2p\pi^+$ orbital and that in the $2p\sigma$ orbital is also random. The spin quantization of the electron in $2p\sigma$ is referred to that of the electron in the hydrogen 1s orbital and is independent of that of the remaining electrons of the oxygen atom.

§ 7. The Valence State of the Oxygen Atom

Let us now examine the state of the oxygen atom in the ground state of the OH radical, as described by the method of § 5. In general, the approximate eigenfunction is given by

$$\Omega_N = \frac{1}{\sqrt{5}}\left\{\mathscr{P}_1 - \sum_{i=1}^{4}\mathscr{P}_{is}\right\}\left[\frac{1}{\sqrt{3}}\Gamma_{PN}\{\Phi(^3P,1,0)\,\bar{h}(5) - \sqrt{2}\Phi(^3P,1,1)\,\bar{h}(5)\}\right.$$
$$\left. + \Gamma_{DN}\Phi(^1D,1,0)\,h(5)\right]. \qquad \ldots\ldots(7.1)$$

The Γ's are continuously varying functions of the internuclear distance r, determined by minimizing the energy. Since the normalization condition

$$\int \Omega_N^* \Omega_N\, dV = C,$$

where C is arbitrary, imposes one condition on these coefficients, it is really only their ratio which is determined by the variational principle. It is usually the custom to set $C=1$; for our purpose, however, it will be more convenient not to normalize to unity in this way, but instead to impose the scale factor in such a way that

$$|\Gamma_{PN}|^2 + |\Gamma_{DN}|^2 = 1,$$

so that, without loss of generality, we may set

$$\Gamma_{PN} = -\cos\lambda, \qquad \Gamma_{DN} = \sin\lambda.$$

It may be noted that λ is defined as a function of r.

In order to describe the oxygen atom as it may be supposed to exist in the OH radical for given r, we now consider a fictional, non-adiabatic dissociation process. We imagine that the atoms be infinitely separated in such a way that λ retains that value which calculation showed to be the best value at the internuclear distance r. (In an adiabatic dissociation process, λ would approach the value 0 as $r \to \infty$.) Our purpose in considering such an artificial atomic separation is to preserve at the oxygen atom that electronic structure with which it may be associated when it is chemically bonded to the hydrogen atom. At infinite internuclear distance the appropriate wave function is then still (7.1). Now, for such distances, electron exchange is infinitely improbable; since there is no interaction between the atoms,

an infinite time will elapse before such exchange may be supposed to occur. By observing that the oxygen atom is electrically neutral—e.g. by noticing that it is not deflected in an electric field—we may label the electrons at this atom serially from I to IV. It is not suggested that these electrons may in any way be distinguished from any other electrons, but rather that in the absence of exchange forces between the atoms no loss of generality occurs by such simplification of our notation. The dissociation products of our ideal process may then be described by the wave function

$$\frac{1}{\sqrt{3}}\cos\lambda\,\Phi(^3\mathrm{P},\,1,\,0)\,\mathrm{h(V)}+\frac{\sqrt{2}}{\sqrt{3}}\cos\lambda\,\Phi(^3\mathrm{P},\,1,\,1)\,\bar{\mathrm{h}}(\mathrm{V})+\sin\lambda\,\Phi(^1\mathrm{D},\,1,\,0)\,\mathrm{h(V)}, \qquad \ldots\ldots(7.2)$$

which is obtained by omitting the first (antisymmetrizing) factor from (7.1) and identifying the electronic coordinates of the oxygen atom by indices I to IV, and those of the hydrogen atom by V. It may be noted that when the atomic functions are normalized to unity, as we suppose to be the case, then (7.2) is itself normalized to unity for these very large values of r.

Now let us examine the statistical interpretation of the wave function (7.2). The probability **P** of finding the oxygen atom in the state $\Phi(^1\mathrm{D},\,1,\,0)$ is simply the sum of the probabilities of finding the oxygen atom in this state and of finding the hydrogen atom in any of its stationary states:

$$\mathbf{P}(^1\mathrm{D},\,1,\,0)=\mathbf{P}\{(^1\mathrm{D},\,1,\,0)_a(\mathrm{h}\bar{\mathrm{h}})\}=|\sin\lambda|^2+0=\sin^2\lambda.$$

Similarly, it is found that

$$\mathbf{P}(^3\mathrm{P},\,1,\,0)=\mathbf{P}\{(^3\mathrm{P},\,1,\,0)_a(\mathrm{h}\bar{\mathrm{h}})\}=\left|-\frac{1}{\sqrt{3}}\cos\lambda\right|^2+0=\tfrac{1}{3}\cos^2\lambda,$$

$$\mathbf{P}(^3\mathrm{P},\,1,\,1)=\mathbf{P}\{(^3\mathrm{P},\,1,\,1)_a(\mathrm{h}\bar{\mathrm{h}})\}=0+\left|\frac{\sqrt{2}}{\sqrt{3}}\cos\lambda\right|^2=\tfrac{2}{3}\cos^2\lambda.$$

It may be noticed that an exactly similar description of the oxygen atom would be given by the function

$$\frac{1}{\sqrt{3}}\cos\lambda\,\Phi(^3\mathrm{P},\,1,\,0)+\frac{\sqrt{2}}{\sqrt{3}}\cos\lambda\,\Phi(^3\mathrm{P},\,1,\,1)+\sin\lambda\,\Phi(^1\mathrm{D},\,1,\,0). \qquad \ldots\ldots(7.3)$$

Apart from the phases of the individual Φ, which are in any case arbitrary, it is clear that (7.1) determines (7.3) uniquely. (We may adopt the convention of setting real and positive coefficients for the atomic functions. This convention is only invalidated in the presence of magnetic fields, which we do not consider here.) In fact, we shall use (7.3) to define the 'state' of the oxygen atom in the OH radical as given by the perturbation method of § 5. In particular, 'the valence state' of the oxygen atom in the ground state of the OH radical is defined by

$$\tfrac{1}{3}\Phi(^3\mathrm{P},\,1,\,0)+\frac{1}{\sqrt{2}}\Phi(^3\mathrm{P},\,1,\,1)+\tfrac{1}{6}\Phi(^1\mathrm{D},\,1,\,0). \qquad \ldots\ldots(7.4)$$

This is obtained by comparing (5.5) with (6.3) and noting that the VB function Θ_{B} corresponds to setting $\lambda=\pi/6$. Our domains (1) to (3) of the figure may, in this notation, be described as those ranges of λ for which $\lambda\simeq 0$, $0<\lambda<\pi/6$ and $\lambda\simeq\pi/6$ respectively.

It will be convenient at this stage to digress a little, and to describe the specification of atomic states. Suppose we have a free oxygen atom. In the absence of any further knowledge, this is described quite generally by a function of the form:

$$\sum_i \Phi^i G^i, \qquad \ldots\ldots(7.5)$$

where Φ^ξ represents a normalized atomic eigenfunction corresponding to a state of energy eigenvalue w^ξ, and the coefficients G^ξ are unknown functions whose arguments depend on the time t. Such a condition of the oxygen atom is said to be non-stationary in so far as its energy is not precisely determined. If we were to make a measurement of the energy we should, by repeating our experiment many times, obtain various values whose mean would be given by

$$\sum_\xi |G^\xi|^2 w^\xi \Big/ \sum_\xi |G^\xi|^2. \qquad \ldots\ldots(7.6)$$

This is called the expectation value of the energy for the non-stationary (or complex) atomic state. On the other hand, if as the result of a particular measurement we find the value w^ξ, we know that immediately following our determination the state of the atomic system is given by the energy eigenfunction $\Sigma_\xi{}^\dagger \Phi^\xi G^\xi$, which describes a so-called stationary state. The summation $\Sigma^\dagger$ is taken over only those independent eigenfunctions Φ^ξ belonging to the same eigenvalue w^ξ, thereby allowing for possible degeneracy. The probability factor associated with this result of performing the energy measurement is simply $\Sigma_\xi{}^\dagger |G^\xi|^2$.

The interpretation to be put upon our definition (7.3) of the state of the oxygen atom in the OH radical is now clear. It represents a complex or non-stationary state of the oxygen atom. Similarly the valence state (7.4) of the oxygen atom is seen to be a particular non-stationary state. To supplement this description of the valence state, we may also use the terminology introduced at the end of the preceding section: it is that non-stationary state which arises from the atomic configuration $(1s)^2(2s)^2(2p\pi^+)^2(2p\pi^-)(2p\sigma)$ and for which the mutual orientation of the spin vectors associated with the electrons in the $2p\pi^-$ and $2p\sigma$ orbitals is random. In the state $\Phi(^1D, 1, 0)$ the orbital picture would describe these spins as being antiparallel, whereas in the ^3P state with $L_z = 1$ they would be said to be parallel. That the epithet 'random' is well chosen is easily seen by noting that the probability of finding it in the ^1D state of $L_z = 1$ is $\frac{1}{3}$. Since there are three ^3P components with $L_z = 1$, and only one ^1D component, the 3 to 1 weighting of these states in (7.4) is indeed apposite.

If, instead of taking the $\Lambda = 1$, $S_z = \frac{1}{2}$ component of the state for discussion, we had used any of the other three components of OH, $^3\Pi$, we should have described the valence state of the oxygen atom by means of one or other of the functions

$$\begin{aligned}
&\tfrac{1}{3}\Phi(^3P, 1, 0) + \frac{1}{\sqrt{2}}\Phi(^3P, 1, -1) + \tfrac{1}{3}\Phi(^1D, 1, 0),\\
&\tfrac{1}{3}\Phi(^3P, -1, 0) + \frac{1}{\sqrt{2}}\Phi(^3P, -1, 1) + \tfrac{1}{3}\Phi(^1D, -1, 0),\\
&\tfrac{1}{3}\Phi(^3P, -1, 0) + \frac{1}{\sqrt{2}}\Phi(^3P, -1, -1) + \tfrac{1}{3}\Phi(^1D, -1, 0). \qquad \ldots\ldots(7.7)
\end{aligned}$$

All these describe valence states of the oxygen atom as non-stationary states in which the statistical weights of the ^3P and ^1D states are $\frac{2}{3}$ and $\frac{1}{3}$ respectively. We may say that these are all components of one and the same valence state, as in any simple superposition of two or more of them.

The expectation value for the energy of any of the conditions (7.4), (7.7) is clearly

$$w_V = \tfrac{2}{3}w_P + \tfrac{1}{3}w_D, \qquad \ldots\ldots(7.8)$$

where w_P, w_D are the energies of the ^3P, ^1D states respectively. w_V, so defined, is called the energy of the valence state. When this energy is referred to the energy

of the ground ^{3}P state of the oxygen atom it is called the promotional energy of the valence state, which is therefore $w_V - w_P = \frac{1}{2}(w_D - w_P)$.

It may be shown that the state of the oxygen atom in the water molecule, as given by the electron-pairing approximation, is in this same valence state: the orbital function representing the simple electron-pairing description of the two OH bonds (assumed, for the moment, to lie along the x and y axes respectively) is

$$\tfrac{1}{2}\{|x\bar{x}y\bar{z}h\bar{h}'| - |x\bar{x}\bar{y}z\bar{h}h'| - |x\bar{x}y\bar{z}h\bar{h}'| + |x\bar{x}\bar{y}zh\bar{h}'|\}. \quad \ldots\ldots(7.9)$$

Here x, y denote those linear combinations of π^+, π^- which have vanishing components of angular momentum about the x and y axes respectively and, for consistency, σ has been replaced by z; h, h′ represent the 1s orbitals of the hydrogen atoms on the z, y axes.

$$x = \frac{1}{\sqrt{2}}(\pi^+ + \pi^-),\ y = \frac{1}{i\sqrt{2}}(\pi^+ - \pi^-),\ z = \sigma.$$

By applying the type of analysis which we have already supplied above, it is easy to see that the valence state of the oxygen atom in the water molecule is given by the function

$$\frac{1}{2\sqrt{2}} \sum_{L_z=\pm 1} \{\Phi(^1\mathrm{D}, L_z, 0) + \sum_{S_z=0,\,\pm 1} \Phi(^3\mathrm{P}, L_z, S_z)\}. \quad \ldots\ldots(7.10)$$

This may also be obtained by simply superposing the four functions (7.4), (7.7) and renormalizing—proving our contention. In this case it is clear that our valence state of the oxygen atom enables it to form two electron-pair bonds by means of its 2p orbitals: it is therefore known as the divalent state of the oxygen atom. Moreover, it is noteworthy that when it only forms one electron-pair bond, as in the OH radical, the oxygen atom is already in its divalent state. Since the spins of the electrons in the singly occupied 2p orbitals must be uncoupled before the first OH bond is formed, the remaining electronic spin is then already free to couple with the spin of the electron on the second hydrogen atom.

§8. Valence State Tables

The valence state concept was first formalized by Van Vleck (1933, 1934) with particular reference to the quadrivalent state of carbon. It was perhaps already implicit in the earlier work of Slater (1931). Mulliken catalogued many valence states of the lighter elements in 1934 for the establishment of his electro-affinity scale (see §11), and gave formulae for the calculation of promotional energies. However, the work of all these authors differs somewhat from the treatment particularized in the preceding sections. We have chosen to emphasize the role of atomic states rather than that of atomic orbitals for reasons which have been given elsewhere (Moffitt 1951 a, b). Accordingly we have defined the energies of valence states in such a way that these may be calculated immediately and uniquely whenever the spectral location of the stationary states of a given atom is known. Since this was not the procedure adopted by Mulliken, we now give tables enabling this computation to be performed for other cases than the oxygen atom.

It will be appreciated how valence states may be defined by means of relations analogous to (7.4) for other atoms in other molecules, always provided that the chemical binding in the latter is attributed to electron-pair bonds. The general case is exemplified by a formula of the type (7.5), where the coefficients G^{ξ} are numerically determined by valence considerations of the kind that led to (7.4) for the OH radical and to (7.10) for the water molecule from the appropriate VB

functions (6.1), (7.9). When these coefficients are determined, the energies of these valence states follow immediately from the simple formula (7.6).

In order to effect a suitable tabulation, we follow Mulliken (1934) and consider only those valence states which may be said to arise from configurations of the type $s^m p^n$, where m and n are positive integers. For linear molecules the appropriate mode of quantization for the 2p orbitals will be the σ, π system. In other cases it is more convenient to use the x, y, z notation. A given valence state is specified at the beginning and end of each row in table 3 by the orbital assignment of its electrons. Whenever a given orbital is only singly occupied it will mean that the spin of the electron associated with it is quantized independently of any other electronic spins of the same atom: it therefore represents a valence electron. The total number of singly occupied orbitals gives the valence number ν of the state in question, which is therefore also labelled V_ν.

The first set of columns in table 3 gives the contribution from each stationary state to the valence state; equivalently, it gives the probability that a non-stationary state of this type be found to have an energy equal to that of the specified stationary state. These figures enable us to calculate the valence state energies immediately. Thus, under $(s^2)\sigma\pi^2\pi$, V_3 we find the factor $\frac{3}{4}$ in the ^3P column and $\frac{1}{4}$ in the ^1D column; we therefore know that the energy of this valence state is given by (7.8). The second set of columns gives the contributions to the energies of these states of certain parameters first introduced by Slater (1929), which formed the basis of Van Vleck and Mulliken's estimates of promotional energies. They are given for completeness, but may be ignored for present purposes.

There is another set of valence states which does not appear in the above tabulation. Such states arise when the atomic orbitals involved in bond formation are of hybrid type. Their discussion involves no additional principles and is easily pursued by simple extension of those developed above. This topic will be resumed rather briefly in § 10 below.

§ 9. Excitation and Bond Energies

In the remaining sections certain applications of the valence state concept will be described by means of simple examples. Of these, the relation between bond dissociation energies and promotional energies is perhaps the most important. Also, it was in this context that the atomic valence state was first introduced by Van Vleck in 1934.

In thermochemical discussions which involve the bond additivity assumption, within the framework of the electron-pairing theory of chemical bonding, it is often found expedient to associate an energy with the valence state of, for example, the oxygen atom. As we have seen above, the valence states of the oxygen atom in the ground states of the OH radical and of the water molecule are the same. The expectation value for the energy of this non-stationary state is greater than that of the ground ^3P state—for it involves some admixture from the ^1D state. This increase in energy is the promotional energy of the valence state, namely $(w_V - w_P)$. It may be described as the energy which is required to uncouple the spins of the electrons in the two singly occupied 2p orbitals of the oxygen atom. Moreover, this uncoupling process is complete after the formation of the first OH bond. Clearly, therefore, we should have to allow for this promotional energy in attempting to predict, or to understand, the relationship between the first and second OH bond dissociation energies of the water molecule. For this purpose it is found most

Table 3. Valence State Tables

$(s^2)p^2$	^3P	^1D	^1S	G_2	$(s^2)p^4$
^3P	1	0	0	0	^3P
^1D	0	1	0	6	^1D
^1S	0	0	1	15	^1S
xy, $\sigma\pi$, V_2	$\frac{3}{4}$	$\frac{1}{2}$	0	$1\frac{1}{2}$	x^2yz, $\pi^2n\sigma$, V_2
x^2, σ^2, V_0	0	$\frac{2}{3}$	$\frac{1}{3}$	9	x^2y^2, $\pi^2\pi^2$, V_0
π^2, V_0	0	1	0	6	$\sigma^2\pi^2$, V_0
$\pi\pi$, V_2	$\frac{3}{4}$	$\frac{1}{12}$	$\frac{1}{6}$	3	$\sigma^2\pi\pi$, V_2

$(s^2)p^3$	$^4S^o$	$^2D^o$	$^2P^o$	G_2
$^4S^o$	1	0	0	0
$^2D^o$	0	1	0	9
$^2P^o$	0	0	1	15
xyz, V_3	$\frac{1}{2}$	$\frac{1}{2}$	0	$4\frac{1}{2}$
$\pi n\sigma$, V_3	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	6
x^2y, $\sigma^2\pi$, $\pi^2\pi$, V_1	0	$\frac{1}{2}$	$\frac{1}{2}$	12
$\pi^2\sigma$, V_1	0	1	0	9

sp	$^3P^o$	$^1P^o$	G_1	sp^5
$^3P^o$	1	0	0	$^3P^o$
$^1P^o$	0	1	2	$^1P^o$
sx, $s\sigma$, $s\pi$, V_2	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	sxy^2z^2, $s\sigma\pi^2\pi^2$, $s\sigma^2\pi^2\pi$, V_2

sp^2	^4P	^3P	^2D	^2S	G_1	G_2	sp^4
^4P	1	0	0	0	0	0	^4P
^2P	0	1	0	0	3	0	^2P
^2D	0	0	1	0	1	6	^2D
^2S	0	0	0	1	1	15	^2S
$s\pi^2$, V_1	0	0	1	0	1	6	$s\sigma^2\pi^2$, V_1
sxy, $s\sigma\pi$, V_2	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	0	1	$1\frac{1}{2}$	sx^2yz, $s\pi^2n\sigma$, V_2
$s\pi\pi$, V_2	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{6}$	1	3	$s\sigma^2\pi\pi$, V_2
$s\sigma^2$, sx^2, V_1	0	0	$\frac{2}{3}$	$\frac{1}{3}$	1	9	$s\pi^2\pi^2$, sx^2y^2, V_1

sp^3	$^5S^o$	$^3S^o$	$^3D^o$	$^1D^o$	$^3P^o$	$^1P^o$	G_1	G_2
$^5S^o$	1	0	0	0	0	0	0	0
$^3S^o$	0	1	0	0	0	0	4	0
$^3D^o$	0	0	1	0	0	0	1	9
$^1D^o$	0	0	0	1	0	0	3	9
$^3P^o$	0	0	0	0	1	0	1	15
$^1P^o$	0	0	0	0	0	1	3	15
$sxyz$, V_4	$\frac{5}{16}$	$\frac{3}{16}$	$\frac{3}{8}$	$\frac{1}{8}$	0	0	$1\frac{1}{2}$	$4\frac{1}{2}$
$sx\pi\sigma$, V_4	$\frac{5}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{9}{16}$	$\frac{1}{16}$	$1\frac{1}{2}$	6
sx^2y, $so^2\pi$, $sn^2\pi$ } V_2	0	0	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	$1\frac{1}{2}$	12
$sn^2\sigma$, V_2	0	0	$\frac{3}{4}$	$\frac{1}{4}$	0	0	$1\frac{1}{2}$	9

convenient to analyse the relevant thermochemistry in terms of the fictional dissociation process which we have used above, and which leaves the oxygen atom in its non-stationary valence state, rather than in its ground state. In this way we should expect to be able to compare the first and second dissociation energies of the OH bonds more profitably.

We therefore make the additivity assumption in the form : the energy of forming an electron-pair bond by coupling the spin of an electron of a hydrogen atom together with that of an electron in an appropriately directed oxygen 2p orbital, whose spin has been uncoupled from those of the remaining oxygen electrons, is a constant. This assumption leads us to identify the OH bond energy $E(\mathrm{O-H})$ with the first OH bond dissociation energy of the water molecule, D_1 say. For both before and after this dissociation process the oxygen valence state remains unaltered. We should then expect the following relation to hold between D_1, $w_V - w_P$ and the adiabatic dissociation energy of the OH radical, D_2 say:

$$D_1 + D_2 \simeq 2E(\mathrm{O-H}) - (w_V - w_P),$$

or, since $D_1 = E(\mathrm{O-H})$,

$$D_2 \simeq D_1 - (w_V - w_P). \qquad \ldots\ldots (9.1)$$

We therefore predict that the second OH bond dissociation energy is lower than the first by an amount equal to the promotional energy of the divalent state of the oxygen atom. (This argument, or a somewhat similar one, was first applied to the H_2O system by Skinner in 1949. Pauling (1949) used the observed values of D_1, D_2 in order to obtain an estimate of $w_V - w_P$ using this relation. He also used four other methods in order to assess this quantity—all of which really correspond to different definitions of the oxygen valence state. Since the values he obtained ranged from 11 to 20 kcal, it is felt that if the valence state concept is to be used at all, a method of definition should be adopted which does not give divergent values, but which leads to a unique figure.)

Let us therefore evaluate the promotional energy of the oxygen valence state and compare D_1 and D_2 on this basis. It may be noted, in passing, that $w_V - w_P$ also determines the asymptotic separation of the E_N, E_B curves of the figure. Using (7.8), we see that the promotional energy is simply one quarter of the observed separation between the ^1D and ^3P states of the oxygen atom, namely $w_V - w_P = 0{\cdot}49$ ev $= 11{\cdot}3$ kcal. Now it is well known that $D_1 = 118$ kcal and that $D_2 = 101$ kcal, so that the difference between the first and second OH bond dissociation energies in the water molecule is rather larger than would have been expected from (9.1). On the other hand, the greater part of the disparity between D_1 and D_2 is to be attributed to the energy of promotion at the oxygen atom, and not to inadequacies of the bond additivity assumption.

The application of similar principles to the methane molecule has been examined in some detail by Voge (1936, 1948).

Another nice and simple illustration of how the bond additivity assumption may be used in conjunction with the valence state idea concerns the lower states of the NH radical.

We first illustrate the conditions under which the NH bond is formed according to the electron-pairing theory: suppose we have a nitrogen atom with electronic configuration $(1s)^2(2s)^2(2p)^3$ and a ground state hydrogen atom. Forming the usual valence bond between h and $2p\sigma$, we have the picture $(1s)^2(2s)^2(2p\pi)^3(2p\sigma \longleftrightarrow h)^2$. Now this does not define any one state of the NH

radical uniquely, owing to the different ways in which two electrons may be assigned to the doubly degenerate $2p\pi$ orbitals. And, in fact, it is easily seen that three distinct states arise:

$$\mathrm{NH}, {}^3\Sigma^- : (1s)^2(2s)^2(2p\pi^+)(2p\pi^-)(2p\sigma \longleftrightarrow \mathrm{h})^2, \quad \ldots\ldots(9.2)$$

$$\mathrm{NH}, {}^1\Delta : (1s)^2(2s)^2(2p\pi^{\pm})^2(2p\sigma \longleftrightarrow \mathrm{h})^2, \quad \ldots\ldots(9.3)$$

$$\mathrm{NH}, {}^1\Sigma^+ : (1s)^2(2s)^2(2p\pi^+)(2p\pi^-)(2p\sigma \longleftrightarrow \mathrm{h})^2. \quad \ldots\ldots(9.4)$$

In the first allocation the two $2p\pi$ electrons are in different components of the doubly degenerate orbital, and their spins are parallel; in the last, they are anti-parallel. Since in both cases there is no resultant angular momentum about the internuclear axis, the respective states are ${}^3\Sigma$ and ${}^1\Sigma$. If, on the other hand, both electrons are considered as filling either the $2p\pi^+$ or the $2p\pi^-$ orbital, we obtain a singlet state for which $\Lambda = \pm 2$, and therefore the doubly degenerate ${}^1\Delta$ state.

Now let us see how these states arise from the more general perturbation theory which we applied to the OH radical in § 5. As is seen from table 3, the ground configuration of the nitrogen atom gives rise to three stationary states, namely ${}^4\mathrm{S}$, ${}^2\mathrm{D}$ and ${}^2\mathrm{P}$. It is easily verified that a ${}^3\Sigma^-$ state of the NH radical can only be formed by bringing together a 1s hydrogen atom and either a ${}^4\mathrm{S}$ or a ${}^2\mathrm{D}$ nitrogen atom. The relevant zeroth order (composite) functions for the $S_z = +1$ component are Ψ'_{S} and Ψ'_{D}; these are defined by means of relations analogous to (5.1)–(5.3):

$$\Psi'_{\mathrm{D}}({}^3\Sigma^-, 0, 1) = \mathscr{A}\Phi({}^2\mathrm{D}, 0, \tfrac{1}{2})\mathrm{h}, \quad \ldots\ldots(9.5)$$

$$\Psi'_{\mathrm{S}}({}^3\Sigma^-, 0, 1) = \tfrac{1}{2}(\sqrt{3}\Psi'_1 - \Psi'_2), \quad \ldots\ldots(9.6)$$

$$\Psi'_1 = \mathscr{A}\Phi({}^4S, 0, \tfrac{3}{2})\bar{\mathrm{h}}, \quad \Psi'_2 = \mathscr{A}\Phi({}^4\mathrm{S}, 0, \tfrac{1}{2})\mathrm{h}, \quad \ldots\ldots(9.7)$$

where $\mathscr{A}$ is an appropriate antisymmetrizer. On the other hand, the ${}^1\Delta$ state can only arise from the ${}^2\mathrm{D}$ state of the nitrogen atom,

$$\Psi({}^1\Delta, \pm 2, 0) = \frac{1}{\sqrt{2}}\mathscr{A}\{\Phi({}^2\mathrm{D}, \pm 2, \tfrac{1}{2})\bar{\mathrm{h}} - \Phi({}^2\mathrm{D}, \pm 2, -\tfrac{1}{2})\mathrm{h}\}, \quad \ldots\ldots(9.8)$$

and the ${}^1\Sigma^+$ state is only formed from ${}^2\mathrm{P}$,

$$\Psi({}^1\Sigma^+, 0, 0) = \frac{1}{\sqrt{2}}\mathscr{A}\{\Phi({}^2\mathrm{P}, 0, \tfrac{1}{2})\bar{\mathrm{h}} - \Phi({}^2\mathrm{P}, 0, -\tfrac{1}{2})\mathrm{h}\}. \quad \ldots\ldots(9.9)$$

Hence, to this degree of approximation, the molecular eigenfunctions of the ${}^1\Delta$ and ${}^1\Sigma^+$ states of the NH radical are uniquely determined by (9.8) and by (9.9). It is easily verified that these are correctly illustrated by the orbital functions associated with (9.3) and (9.4). The ${}^3\Sigma^-$ state, however, is in general represented by a linear combination of the two composite functions (9.5) and (9.6):

$$\Omega({}^3\Sigma^-, 0, 1) = \Psi'_{\mathrm{S}}\Gamma_{\mathrm{S}} + \Psi'_{\mathrm{D}}\Gamma_{\mathrm{D}}. \quad \ldots\ldots(9.10)$$

As the internuclear distance r increases indefinitely, the molecule in its ${}^3\Sigma^-$ state dissociates adiabatically into a ground state hydrogenatom and the ${}^4\mathrm{S}$ nitrogen state —i.e. Γ_{D} vanishes asymptotically. Moving in from this domain of S, L coupling, we pass into the region of transition, where Ψ'_{S}, Ψ'_{D} begin to interact. Finally, in the neighbourhood of the equilibrium internuclear separation we enter the domain of V coupling, according to the electron-pairing theory. For the ${}^1\Delta$ and ${}^1\Sigma^+$ states, however, there is no distinction between these zones: the domains of S, L and of V coupling accidentally coincide and cover the whole range.

It only remains to evaluate Γ_{S}, Γ_{D} of (9.10) on the basis of the VB description (9.2) of the ${}^3\Sigma^-$ state. We notice that the latter may be written as

$$\tilde{\Theta}_{\mathrm{B}} = \frac{1}{\sqrt{2}}\{|\pi^+\pi^-\sigma\bar{\mathrm{h}}| - |\pi^+\pi^-\bar{\sigma}\mathrm{h}|\}. \quad \ldots\ldots(9.11)$$

By expressing this in terms of orbital approximations to the nitrogen atomic eigenfunctions, and hence as approximate composite functions, it is readily verified that Θ_B of (9.11) illustrates the VB function

$$\Theta_B = \frac{1}{\sqrt{3}}\{\sqrt{2}.\Psi_S - \Psi_D\}. \qquad \ldots\ldots(9.12)$$

In exactly the same way (6.1) illustrated (6.3) for the OH radical. Thus, in the domain of V coupling, Γ_S, Γ_D assume the values $\sqrt{2}/\sqrt{3}$, $-1/\sqrt{3}$ respectively, and the energy of the concomitant nitrogen valence state, referred to its ^4S ground state, is one third of the separation between its ^2D and ^4S states.

The valence states of the nitrogen atom in these three molecular states may be described as follows. For both $^3\Sigma^-$ and $^1\Sigma^+$ states the appropriate nitrogen atom states arise from the spatial configuration $\pi^+\pi^-\sigma$ of its three 2p electrons; in both, the spin of the electron in the $2p\sigma$ orbital is quantized independently of that of the $2p\pi$ electrons, representing, as it does, the valence electron of the atom. In the $^3\Sigma^-$ states the two remaining spins are parallel, whereas in the $^1\Sigma^+$ state these are antiparallel. The probabilities of finding the valence state apposite to $^3\Sigma^-$ in the ^4S, ^2D states are $\frac{2}{3}$, $\frac{1}{3}$ respectively; the valence state corresponding to $^1\Sigma^+$ may be identified with a component of the ^2P state. Neither valence state is listed in table 3. For the $^1\Delta$ state of the molecule the nitrogen valence state $\pi^2\sigma$ may be identified with a component of the ^2D state.

Let us now apply the bond additivity assumption. According to this, the energy liberated in pairing together the spin of a hydrogen 1s electron and that of an electron in a nitrogen $2p\sigma$ orbital, whose spin has been uncoupled from the remaining spins on this atom, is a constant function of the internuclear separation. The assumption in this case has a rare predictive immediacy. For it tells us that the potential energy curves representing the $^1\Sigma^+$ and $^1\Delta$ states, whose nitrogen valence states are (accidentally) stationary states, should be identical in every way, but that the former should be a distance equal to the separation of the ^2P and ^2D states (1·2 ev) above the latter. This expectation is remarkably well substantiated by the spectral information on the NH radical. Thus the equilibrium internuclear distances in its $^1\Sigma^+$ and $^1\Delta$ states are nearly identical and the separation of their ground states here is found to be 1·1 ev (Herzberg 1950). The potential energy curve of the $^3\Sigma^-$ state should be the same as those of the $^1\Sigma^+$ and $^1\Delta$ states, at least within its domain of coupling, except that it should be displaced vertically downwards in such a way that the separations between the three curves are only determined by the promotional energies associated with their nitrogen valence states. In confirmation of this, the equilibrium internuclear distance for the $^3\Sigma^-$ state is exceedingly close to those of the other two. The relative location of singlet and triplet states of the NH radical is not known. However, since the ^2D–^4S separation of the nitrogen atom is known to be 2·4 ev, it is easy to predict that the ground $^3\Sigma^-$ state of the NH radical should lie some 1·6 ev below the $^1\Delta$ state and 2·8 ev below the $^1\Sigma^+$ state in the neighbourhood of their equilibrium internuclear distances.

§ 10. Hybridization

There are two additional factors which are important in the understanding of the primary sources of chemical binding which have not been considered so far. One concerns orbital hybridization and the other, which will be examined in § 11, accounts for the heteropolar character of certain bonds. In the next few paragraphs we shall discuss some features involved in hybridization.

One of the main reasons for the initial success of the electron-pairing theory was the realization by Slater (1931) and by Pauling (1931) that the valence orbital of a particular atom need not be purely p or s in character, but could assume in part the character of both. It was apparent that a criterion for the formation of strong electron-pair bonds was the extent to which the atomic orbitals concerned were capable of mutual overlap. The more strongly a particular orbital was directed at its partner, the stronger the resultant bond. Now a consideration of the form of s and p functions shows that an admixture of both has better directional qualities than either separately possesses (see, for example, the electron density contours in Moffitt and Coulson 1947). The conclusions which may be drawn from this observation are well known, particularly for quadrivalent carbon. We shall illustrate certain simple factors relating to atomic valence states which are the concomitants of orbital hybridization, using the HCl molecule as an example.

Let us suppose, quite generally, that the chlorine atomic orbital which is responsible for the electron-pair bond with a hydrogen atom is defined by a hybridization parameter χ:

$$\phi(3t\sigma) = \cos\chi\phi(3p\sigma) + \sin\chi\phi(3s),$$

where the symbol ϕ defines the atomic orbitals concerned. There will be six other electrons of the M shell associated with the chlorine atom all of which must be assigned to orbitals orthogonal to $3t\sigma$. Four will complete the $3p\pi$ shell and the remaining lone pair will be in that s–pσ hybrid orbital which is orthogonal to $3t\sigma$, namely

$$\phi(3u\sigma) = -\sin\chi\phi(3p\sigma) + \cos\chi\phi(3s).$$

When $\chi = 0$ there will be no hybridization, and the HCl bond will involve only p character from the chlorine atom. By increasing χ from 0 to about $\pi/4$, the strength of the bond increases progressively (Förster 1939, Moffitt 1950b).

In any case, the orbital description of the molecule may be symbolized by $(u\sigma)^2(t\sigma\leftarrow\rightarrow h)^2$; the associated orbital function is

$$\widetilde{\Theta}(\chi) = \frac{1}{\sqrt{2}}\{|\,u\bar{u}t\bar{h}\,| - |\,u\bar{u}\bar{t}h\,|\},$$

where, for convenience, we have omitted the K, L and $3p\pi$ electrons, all of which are in closed shells or sub-shells. This may be expressed in terms of the non-hybridized functions as follows:

$$\widetilde{\Theta}(\chi) = \frac{1}{\sqrt{2}}\cos\chi\{|\,s\bar{s}\sigma\bar{h}\,| - |\,s\bar{s}\bar{\sigma}h\,|\} + \frac{1}{\sqrt{2}}\sin\chi\{|\,\sigma\bar{\sigma}s\bar{h}\,| - |\,\sigma\bar{\sigma}\bar{s}h\,|\}.$$

It is clear that the coefficient of $\cos\chi$ represents an electron-pair bond involving only the $3p\sigma$ orbital of the chlorine atom, which is therefore in its ground ^2P state $(K)(L)(3s)^2(3p)^5$; the coefficient of $\sin\chi$ on the other hand symbolizes the use of the 3s orbital for this purpose and can only arise from the chlorine atom in its excited ^2S state $(K)(L)(3s)(3p)^6$. $\widetilde{\Theta}(\chi)$ therefore illustrates the VB function

$$\Theta(\chi) =$$

$$\frac{1}{\sqrt{2}}\mathscr{A}[\cos\chi\{\Phi(^2P, 0, \tfrac{1}{2})\bar{h} - \Phi(^2P, 0, -\tfrac{1}{2})h\} + \sin\chi\{\Phi(^2S, 0, \tfrac{1}{2})\bar{h} - \Phi(^2S, 0, -\tfrac{1}{2})h\}],$$

for which the valence state of the chlorine atom is

$$\frac{1}{\sqrt{2}}\cos\chi\{\Phi(^2P, 0, \tfrac{1}{2}) + \Phi(^2P, 0, -\tfrac{1}{2})\} + \frac{1}{\sqrt{2}}\sin\chi\{\Phi(^2S, 0, \tfrac{1}{2}) + \Phi(^2S, 0, -\tfrac{1}{2})\}.$$

The promotional energy of this valence state is simply the product of $\sin^2\chi$ and the excitation energy of the ^2S state of the chlorine atom.

Since the difference in energy between the ^2S and ^2P states of the chlorine atom is large, the tendency for stronger H–Cl binding, by increasing χ, will be partially offset by the energy of promotion at the chlorine atom. The optimum value for χ will be a function of the internuclear distance which must be determined by minimizing the energy of the system.

This possibility of hybridization has certain interesting repercussions, since the physical properties of the HCl molecule are very sensitive to χ. In particular, the electronic density around the chlorine atom is now symmetric and must be associated with an 'atomic' dipole moment (Moffitt 1949, 1950 b, Robinson 1949). Moreover, the electric quadrupole coupling constant of the chlorine nucleus is also affected by these considerations (Townes and Dailey 1949). The specific type of hybridization illustrated here has been called second-order hybridization, since it involves no change in the valence number of the chlorine atom (see Moffitt 1950 a, where hybrid states have been classified, and also Mulliken 1952).

§ 11. Electronegativities

The last subject which we have to consider concerns the contribution of ionic terms, and the establishment of an electronegativity scale. For this purpose we shall be guided by Mulliken's arguments (1934, see also Hund 1931), rather than those of Pauling (1932). Pauling bases his scale on a special interpretation of thermochemical data, the validity of which may be questioned; and the detailed examination of his postulates falls outside the limits of the present review. Mulliken's scale arises quite naturally from the preceding analysis, however, and will be illustrated by means of the hydrogen fluoride molecule.

In addition to the covalent electron-pair bond between hydrogen and fluorine, we admit ionic structures and consider three VB functions, which may be defined in orbital symbols as follows:

$$\Theta(\mathrm{H–F}):\ (1s)^2\,(2s)^2\,(2p\pi)^4\,(2p\sigma\longleftrightarrow h)^2,$$

$$\Theta(\mathrm{H^-F^+}):\ (1s)^2\,(2s)^2\,(2p\pi)^4\,(h)^2,$$

$$\Theta(\mathrm{H^+F^-}):\ (1s)^2\,(2s)^2\,(2p\pi)^4\,(2p\sigma)^2. \qquad \ldots\ldots(11.1)$$

In the first of these structures the two electrons associated with the bond are each assigned to one or other of the two atoms. In the second, both are to be associated with the hydrogen atom, and in the last both have migrated into the $2p\sigma$ orbital of the fluorine atom. The two latter functions are stabilized by primary Coulomb forces. In general, therefore, the molecular eigenfunction for the ground state of the molecule is expressed as a linear combination of these three VB functions:

$$\Omega = \Theta(\mathrm{H-F})\,\Gamma_1 + \Theta(\mathrm{H^-F^+})\,\Gamma_2 + \Theta(\mathrm{H^+F^-})\,\Gamma_3. \quad \ldots\ldots(11.2)$$

(Second-order hybridization at the fluorine atom has been neglected for the moment.) The first task of an electronegativity scale, and the only one which will concern us here, is to establish an inequality between Γ_2 and Γ_3. For if $|\Gamma_3| > |\Gamma_2|$, as is most likely, then we should say that fluorine is more electronegative than hydrogen.

Now it is, of course, always possible to make calculations for every particular molecule which will give us the desired information, but the whole object of the

scale is to predict the result of such calculations on the basis of properties of the isolated atoms from which the molecule is built up. We wish to assign numbers, namely electronegativities, to atoms which will assess their electronegative propensities with little reference to their specific molecular environments.

It may well be questioned whether such a scale can in fact ever be set up—whether it is possible to talk of the electronegativity of an atom without reference to a particular set of internuclear distances in some given state of a specified molecule. However, there is no doubt that it makes sense to say that fluorine is more electronegative than sodium, and even in those cases where the distinction between two atoms is less marked, much has been gained by the use of the electronegativity concept.

We shall therefore postulate that the scale exists, and that it is well defined for ground state problems whenever the valence states of the participating atoms and their ions are given.

This postulation will be amplified in the sequel. But it may be noted straight away that we have introduced it within the framework of the electron-pairing approximation. It is very doubtful whether a theoretical scale can be constructed on the basis of a postulation which transcends this level of generality. Moreover, it has been implied that the electronegativity of an atom is not a constant number, but depends upon its valence condition, both in its neutral and in its ionic states: it is tacitly supposed that electronegativity is a function of an atom's molecular environment in so far—and no farther—as this dictates its valence states. The electronegativity of an atom is determined when its bonding orbitals have been specified.

Returning to our example, we note that the valence state of the fluorine atom in the covalent condition $\Theta(H\text{–}F)$ is to be identified with a component of its ^2P ground state. In its ionic condition $\Theta(H^+F^-)$, the valence state is the ^1S ground state of F^-. When it is positively charged in $\Theta(H^-F^+)$, however, its valence state may be symbolized by $s^2\pi^4$, V_0 which, as is seen from table 3, consists of components from the excited ^1D and ^1S levels of F^+, with statistical weights of $\frac{2}{3}$ and $\frac{1}{3}$ respectively.

Now let us consider the behaviour of the function (11.2) as the internuclear distance between the two atoms is progressively increased from its equilibrium value. The condition for the relation $|\Gamma_3|>|\Gamma_2|$ to hold is then easily shown to become

$$\{W(H^+, F^-) - W(H, F)\}\left|\int \Theta^*(H-F)\,\mathscr{H}\,\Theta(H^-F^+)\,dV\right| < \{W(H^-, F^+) - W(H, F)\}\left|\int \Theta^*(H-F)\,\mathscr{H}\,\Theta(H^+F^-)\,dV\right|, \quad \ldots\ldots(11.3)$$

where $W(H, F)$ is the energy of the isolated hydrogen and fluorine atoms, $W(H^+, F^-)$ the energy of the infinitely separated fluorine ion and proton, and $W(H^-, F^+)$ the expectation energy of the separated negative hydrogen ion and the positively charged fluorine valence state $s^2\pi^4$, V_0. The integrals which appear as factors on both sides of this inequality are rapidly varying functions of the internuclear distance, vanishing asymptotically. They represent the more specific interactions of the ionic with the covalent terms in (11.2). Moreover, their values are primarily dictated by the relative size of the participating atoms and ions; they are functions acutely dependent upon the geometry of the one atom with respect to that of its partner in bonding. If, therefore, electronegativity is to be an absolute property

determined solely by the valence states of an atom and its ions, the inequality (11.3) must derive its validity from the relation

$$W(H^+, F^-) - W(H, F) < W(H^-, F^+) - W(H, F). \quad \ldots\ldots(11.4)$$

The relative difference between the moduli of the interaction integrals must be outweighted by the relative difference symbolized in (11.4).

Mulliken bases his electronegativity scale on this last inequality, which may be written, more revealingly, in the form

$$I_H + E_H < I_F + E_F, \quad \ldots\ldots(11.5)$$

where I and E represent respectively the (valence state) ionization potential and electron affinities of the atoms concerned. Thus I_F is the energy required to remove an electron from the ground state of a fluorine atom, leaving the positive ion in its complex state $s^2\pi^4$, V_0, whose expectation energy lies one third of the way between that of its ^1D and ^1S states. The quantity $I_F + E_F$ gauges additively the difficulty of removing an electron from the fluorine atom's valence orbital, and the ease with which this will accept one. It is therefore a very appealing criterion of electronegativity (for some applications see, for example, Mulliken 1935).

The absolute electronegativity scale which Mulliken constructed in this way, using $I + E$, agrees very well with that of Pauling in those cases where the comparison can be made. It is, however, vital for its success that the relevant ionization potentials and electron affinities be referred to the correct valence states of the atoms.

There is another factor which it is most important to bear in mind when assessing electronegativities in this way. It is generally true that the removal of an s electron from an atom is more difficult than the removal of a p electron. Thus when we wish to determine the electronegativity of an atom which forms bonds by means of hybrid orbitals, the associated ionization processes must be defined in a manner which gives proper attention to this effect—namely, by combining the analysis exemplified in § 10 with that of the present section. This has been discussed by Mulliken (1934) and by Moffitt (1950 a, b). As is to be expected, it emerges that the greater the s character of a given bonding orbital, the greater the electronegativity it shows. The most striking differences of electronegativity exhibited by the same atom naturally occur for carbon, whose valence conditions may prescribe varying types of hybridization. Thus the greater electronegativity of the trigonally hybridized carbon sp^2 orbital over its tetrahedral sp^3 counterpart is shown by the hydrogen exchange which occurs on dissolving metal alkyls in benzene. And the still greater electronegativity of the digonal sp hybrid orbitals is manifested in the acidic properties of acetylene. This may be taken as a demonstration of the necessity of specifying the valence states of atoms before ascribing to them meaningful electronegativities.

References

Condon, E. U., and Shortley, G. H., 1935, *Theory of Atomic Spectra* (Cambridge: University Press).

Förster, T., 1939, *Z. phys. Chem.* B, **43**, 58.

Hartree, D. R., 1946, *Reports on Progress in Physics*, **11**, 113 (London: Physical Society).

Hartree, D. R., and Swirles, B., 1939, *Phil. Trans. Roy. Soc.* A, **238**, 229.

Heitler, W., and London, F., 1927, *Z. Phys.*, **44**, 455.

Heitler, W., and Rumer, G., 1931, *Z. Phys.*, **68**, 12.

Herzberg, G., 1950, *Diatomic Spectra* (New York: Van Nostrand).

HUND, F., 1931, *Z. Phys.*, **73,** 1.
KOTANI, M., 1951, Conference on "Quantum Mechanical Methods in Valence Theory" (Office of Naval Research, Washington D.C.)
LONDON, F., 1930, *Z. phys. Chem.* B, **11,** 222.
MOFFITT, W., 1949, *Proc. Roy. Soc.* A, **196,** 524; 1950 a, *Ibid.*, **202,** 534; 1950 b, *Ibid.*, **202,** 548; 1951 a, *Ibid.*, **210,** 224; 1951 b, *Ibid.*, **210,** 245.
MOFFITT, W., and COULSON, C. A., 1947, *Phil. Mag.*, **38,** 634.
MULLIKEN, R. S., 1934, *J. Chem. Phys.*, **2,** 782; 1935, *Ibid.*, **3,** 573; 1952, *J. Phys. Chem.*, **56,** 295.
NORDHEIM-PÖSCHL, G., 1936, *Ann. Phys., Lpz.*, **20,** 258; **26,** 281.
PAULING, L., 1931, *J. Amer. Chem. Soc.*, **53,** 1367; 1932, *Ibid.*, **54,** 3570; 1933, *J. Chem., Phys.*, **1,** 280; 1949, *Proc. Nat. Acad. Sci., Wash.*, **35,** 229.
ROBINSON, D. Z., 1949, *J. Chem. Phys.*, **17,** 1022.
SKINNER, H. A., 1949, *Trans. Faraday Soc.*, **45,** 20.
SLATER, J. C., 1929, *Phys. Rev.*, **34,** 1293; 1931, *Ibid.*, **37,** 481; **38,** 1109.
TOWNES, C. H., and DAILEY, B. P., 1949, *J. Chem. Phys.*, **17,** 782.
VAN VLECK, J. H., 1933, *J. Chem. Phys.*, **1,** 177, 219; 1934, *Ibid.*, **2,** 20, 297.
VOGE, H. H., 1936, *J. Chem. Phys.*, **4,** 581; 1948, *Ibid.*, **16,** 984.
WIGNER, E., and WITMER, E. E., 1928, *Z. Phys.*, **51,** 859.

ANTIFERROMAGNETISM*

By A. B. LIDIARD†
Department of Physics, University of Illinois, Urbana, Illinois, U.S.A.

CONTENTS

	PAGE
§1. Introduction	202
§2. Earlier studies	203
2.1. Introduction	203
2.2. The theory of Van Vleck	206
2.3. Early experimental studies	210
§3. Later experimental studies	214
3.1. Neutron diffraction	214
3.2. Magnetic anisotropy	219
3.3. Thermal expansion and deformation	222
3.4. Antiferromagnetic resonance	224
§4. Theoretical	225
4.1. Introduction	225
4.2. Extended molecular field calculations	226
4.3. Super-exchange	229
4.4. More accurate calculations	233
§5. The transition metals	237
Acknowledgments	241
References	241

Abstract. The familiar Heisenberg theory of ferromagnetism in the 3d transition metals iron, cobalt and nickel, rests on the assumption of a positive exchange integral between neighbouring ions, since ferromagnetism is not possible in cases where this integral is negative (' antiferromagnetic '). The idea that the remaining non-ferromagnetic members of the 3d series (Sc, Ti, V, Cr, Mn) therefore have negative exchange integrals led to a theoretical study of the properties of antiferromagnetic systems. Néel was the first to show that such a system has a critical temperature, T_0, below which the atomic moments are arranged alternately parallel and antiparallel. Above T_0 the moments are disordered, as in a ferromagnetic above its Curie temperature. The theory was later extended by Van Vleck and it then became apparent that the antiferromagnetic model closely fitted a number of simple *compounds* of the transition metals (e.g. CrSb, MnO, MnF_2). These early developments, which had occurred by the first years of the war, are described in §§1 and 2. Particular attention is given to Van Vleck's theory, as this forms a starting point for much subsequent work. In §3 later experimental studies of antiferromagnetic compounds are reviewed, especial attention being paid to the results obtained by the methods of neutron diffraction. Measurements of magnetic anisotropy and of the slight distortions which occur in many antiferromagnetic lattices on ordering, are also described in detail, and the theoretical significance of the observations is discussed incidentally. The theoretical attempts to improve the Van Vleck theory, both by extending the model and by using more accurate statistics, are considered in §4. This section also includes a discussion of the superexchange mechanism, which accounts for the observation that the strongest interactions in, say, MnO are between next nearest neighbouring Mn ions and not between nearest neighbours as one might expect. The final section (§5) contains a brief survey of the present status of the non-ferromagnetic transition metals themselves; it is shown how the recent neutron diffraction studies by Shull and Wilkinson favour a collective electron description, rather than the sort of description which is successful for the antiferromagnetic compounds. The way to extend the present collective electron theory so that it may give antiferromagnetic ordering has recently been indicated by Slater. The Report concludes with a table listing experimental papers on antiferromagnetism by compound and by the property studied. This should be useful since it is often difficult to trace the experimental work on a particular compound of interest.

* This work was supported in part by the U.S. Office of Naval Research.

† Present address (1954): Department of Physics, University of California, Berkeley, California, U.S.A.

§1. Introduction

In theoretical investigations of the properties of solids, two different approaches are widely used. (For details we refer to the books by Seitz 1940, and by Mott and Gurney 1948.) Both of these attempt to describe the behaviour of a solid in terms of one-electron wave functions. In one, the so called Heitler–London approach, the one-electron functions are taken to be essentially the same as in the free atom, i.e. each one is strictly localized about some particular atom of the solid. In the other, the collective electron approach, the wave functions are non-localized and extend throughout the crystal lattice. In practice neither can be used to give a rigorous theory and the two methods in general must be regarded as complementary. This can be seen particularly clearly in the theory of polar crystals (Mott and Gurney 1948, chap. III). In other systems one method may nevertheless be far more convenient than the other. Thus in discussing the outer valence electrons of metals only the collective description is used, although in the theory of ferromagnetism both have been applied successfully (cf. Stoner 1947).

The subject matter of this Report is divided into two parts, (1) the paramagnetic transition metals of the iron series and (2) a number of compounds of the transition metals (including the ferromagnetic iron, cobalt and nickel) with elements from the last three columns of the periodic table, i.e. from groups V, VI and VII. The properties of the compounds have been (and will be here) described entirely in terms of the Heitler–London theory, whereas for the metals themselves both theories have been employed. For the metals there is still considerable uncertainty how far the predictions of either theory are verified experimentally (§5). Nevertheless it is convenient here to assume the Heitler–London model, in order to introduce the idea of an antiferromagnetic medium.

In the Heisenberg (1928) theory of ferromagnetism, which was the original application of the Heitler–London method to solids, there arises the question of the sign of the exchange integral between two (3d) wave functions centred about neighbouring nuclei. In order that the model be ferromagnetic it is necessary for this integral to be positive, and it is *assumed** that this is so for the three metals iron, cobalt and nickel. But the related paramagnetic transition metals, those from scandium to manganese, must then be described by a negative ('antiferromagnetic') exchange integral. This is the same sign as those encountered in the valence bond theory of molecules (see for example Coulson 1952) and we may therefore expect that the spin moment of an electron on one atom will (at least part of the time) be coupled antiparallel to the moments of those electrons on neighbouring atoms. If the atomic lattice is such that none of the nearest neighbours of one atom are nearest neighbours of one another (e.g. body-centred cubic), we may further expect this antiparallel coupling to spread throughout the crystal, dividing it into two interlocking sub-lattices. Whether such a description is ever applicable to, say, chromium or manganese, is doubtful, but applied for example to the manganese ions in MnF_2 it is directly verified by neutron diffraction experiments (Erickson and Shull 1951). The theoretical problem of antiferromagnetism is to give a detailed theory of the ordering of the spin moments; this must describe how the degree of order changes with temperature, and the effect this has on magnetic and thermal properties, e.g. susceptibility, magnetic anisotropy, specific heat and changes in lattice parameters with temperature.

* Two *a priori* calculations of the integral (Wohlfarth 1949, Kaplan 1952) are in direct contradiction, one yielding a negative, the other a positive result.

In the next section we shall briefly review the studies of antiferromagnetism which had been made by the early years of the war. In §3 we shall survey the experimental work which has been done since that time. Such a survey was made in considerable detail by Bizette in 1946 (see also Bizette 1951, Bizette and Tsai 1949), so that we shall give greater prominence to the knowledge which has been added since then. Similarly §4, on the theoretical research, may be regarded as a supplement to the articles by Van Vleck (1951) and Street (1951). In the final section (§5) we shall briefly summarize the work which has been done in an attempt to settle the controversy over the 'best' description of the electronic structure of the transition metals, the choice being between the collective electron band theory on the one hand and the antiferromagnetic description due to Néel (1936b) and Zener (1951) on the other.

In studying antiferromagnetism it is often very difficult to trace the experimental work on some particular compound of interest. We have therefore included, in addition to the usual list of references, a table giving experimental papers by compound and by the property studied.

There is a great deal in the theory of antiferromagnetism which is also common to other co-operative phenomena, e.g. the order–disorder transformation in alloys, regular solutions and ferromagnetism; for the first two the reader may care to refer to the recent book by Guggenheim (1952), and for the last to the reviews by Stoner (1947) and Van Vleck (1945). At this point we would also like to refer *en bloc* to the many interesting papers on ferromagnetism and antiferromagnetism which were read at the Grenoble and Washington conferences. These were published in the March 1951 issue of the *Journal de Physique et le Radium* and in the January 1953 issue of *Reviews of Modern Physics* respectively. The study of the ferrites may be regarded as applied antiferromagnetism, but in view of the recent Report by Fairweather, Roberts and Welch (1952) we feel that it would be superfluous to make more than a passing reference to this subject here.

§2. Earlier Studies

2.1. *Introduction*

As a starting point for a theory of many paramagnetic substances it has proved a fair approximation to regard the magnetically effective ions as being in S states, i.e. the orbital angular momentum of the 3d electrons is quenched by the crystalline electric field. (For an understanding of how this situation comes about see Van Vleck 1932, 1940 and also Fowler and Guggenheim 1949, especially §1414.) In this picture each ion is regarded as being in a state with a total orbital quantum number $L=0$ and a total spin quantum number S. If the ions are far enough apart for their mutual interactions to be negligible then it is well known that the system is a simple paramagnetic whose induced moment M in the presence of a field H is given by

$$M = NSg\beta B_s\left(\frac{g\beta SH}{kT}\right), \qquad \ldots\ldots(2.1)$$

where β is the Bohr magneton, N is the total number of magnetic ions in the specimen and B_s is the Brillouin function,

$$B_s(y) = \frac{2S+1}{2S}\coth\left(\frac{(2S+1)y}{2S}\right) - \frac{1}{2S}\coth\left(\frac{y}{2S}\right).$$

Strictly the gyromagnetic ratio g should be 2, since we are assuming the absence of any orbital momentum. Nevertheless small amounts of residual orbital moment may be taken into account by allowing g to take on values slightly different from 2. For $\beta H/kT \ll 1$ the relation (2.1) may be simplified and gives for the susceptibility χ,

$$\chi = \frac{M}{H} = \frac{Ng^2\beta^2 S(S+1)}{3kT}. \qquad \ldots\ldots(2.2)$$

This is the well-known Curie law for paramagnetic substances. But when exchange interactions between the ions are present these equations are no longer sufficient and in this section we shall review the earlier studies of the changes which these interactions bring about.

We shall suppose (Dirac 1947, but see also Slater 1953) that the exchange coupling between the ions is described by an equivalent Hamiltonian operator,

$$V^{op} = -\sum_{i,j} J_{ij}(2S^2 + 2\mathbf{S}_i^{op} \cdot \mathbf{S}_j^{op}), \qquad \ldots\ldots(2.3)$$

in which $\mathbf{S}_i^{op}$ is the spin vector operator of ion i and the summation is over all pairs of ions (i, j) for which the exchange integral J_{ij} is not negligible. There is little *a priori* knowledge about the exchange integrals and it was at first assumed that J_{ij} was zero unless ions i and j were nearest neighbours, when it was negative ($\equiv -|J|$). Then dropping the constant term we get,

$$V^{op} = 2|J| \sum_{i,j} \mathbf{S}_i^{op} \cdot \mathbf{S}_j^{op}, \qquad \ldots\ldots(2.4)$$

where the summation is now over all distinct pairs of nearest neighbours.

The question of the behaviour to be expected of a crystal model described by the Hamiltonian (2.3) or (2.4) seems to have been first considered by Bethe (1931), who made an approximate calculation of the ground state energy of a linear chain of atoms. The question was raised again, this time in classical form (replacing the operators $\mathbf{S}_i^{op}$ by classical vectors), by Néel (1932, 1936a), but the characteristic properties of an antiferromagnetic medium were first* brought out by Néel (1936b) using the molecular field approximation. (This approximation is described in greater detail in §2.2.) Néel supposed that the lattice of the magnetic ions could be divided into two equivalent interlocking sub-lattices. The body centred tetragonal lattice shown in fig. 12 is one example; the corner ions belong to one sub-lattice, the central ion belongs to the other. He then found that below a critical temperature T_c there existed an ordered condition in which the spins on one sub-lattice were coupled antiparallel to those on the other. Furthermore the lower the temperature the greater the apparent strength of the coupling. On the other hand above T_c the spins were randomly oriented relative to one another. In other words if $\mathbf{S}_1^{op}$ and $\mathbf{S}_2^{op}$ be the spin vector operators for *any* two ions, the average (expectation) value of $\mathbf{S}_1^{op} \cdot \mathbf{S}_2^{op}$ is zero above the critical temperature, whichever pair of ions we choose. Below T_c the average value $\langle \mathbf{S}_1^{op} \cdot \mathbf{S}_2^{op} \rangle$ is negative if the ions are on different sub-lattices, but positive if they are on the same sub-lattice. This is shown schematically in fig. 1. (We notice that $\langle \mathbf{S}_1^{op} \cdot \mathbf{S}_2^{op} \rangle$

* In his earlier papers Néel calculates the partition function for an antiferromagnet using approximations which would not now be regarded as justified. He assigns too great an importance to 'fluctuations' (see Stoner 1947) and does not obtain a typically co-operative behaviour. His results are nevertheless correct in the limits of very high and of very low temperatures.

is essentially the 'bond order' used by chemists (Penney 1937). Hence since it depends only on whether or not the ions belong to the same sub-lattice and not on their separation, any attempt to describe an antiferromagnet as resonating between different chemical 'structures' (e.g. Harrison 1950) must include all possible 'long bonds'. For further explanations of these terms see Coulson 1952, especially chap. IX.

Néel did not allow for any crystalline anisotropy in his model and since the exchange operator (2.4) is isotropic, the absolute direction of the spin moments is arbitrary in the absence of an applied magnetic field. Such a field, however, defines an axis of symmetry, and below T_c the moments set themselves perpendicular to it ($\langle \mathbf{H} \cdot \mathbf{S}_i^{op} \rangle = 0$). Actually for non-vanishing fields the moments are inclined slightly towards the field direction, giving (in the molecular field approximation) a constant paramagnetic susceptibility. By applying spin-wave theory as it has been formulated by Heller and Kramers (1934), Hulthén (1936)

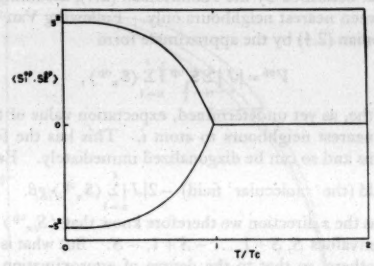


Fig. 1. Schematic diagram showing the temperature variation of the average value $\langle \mathbf{S}_1^{op} \cdot \mathbf{S}_2^{op} \rangle$ as given by 'molecular field' theories. The upper branch applies when atoms 1 and 2 are on different sub-lattices and the lower branch when they are on the same sub-lattice. More accurate theories indicate that $\langle \mathbf{S}_1^{op} \cdot \mathbf{S}_2^{op} \rangle$, although small, is not zero at the Curie point and above—at least when atoms 1 and 2 are nearest neighbours of each other. Furthermore the magnitude of the low temperature limit of $\langle \mathbf{S}_1^{op} \cdot \mathbf{S}_2^{op} \rangle$ will actually be less than S^2 (Kramers 1952).

had previously obtained a formula for the susceptibility which reduced to Néel's expression in the limit $T \to 0$. As the temperature was increased the susceptibility was found to decrease slowly rather than remaining constant as in Néel's theory. However, to return to the molecular field description, Néel found that above T_c, when all long range order has disappeared, the susceptibility decreases according to the law

$$\chi = \frac{Ng^2\beta^2 S(S+1)}{3k(T+\theta)} . \qquad \ldots\ldots(2.5)$$

The constant θ is equal to T_c if ions on one sub-lattice interact only with ions on the other. χ is continuous at T_c. The order–disorder change at T_c is also accompanied by a sharp maximum in the specific heat.

Now in real crystals we may anticipate the existence of anisotropy forces which constrain the spin moments to point only in certain crystallographic directions. Bitter (1938) repeated Néel's calculation for a model with $S = \frac{1}{2}$ and for which the

magnetic moments of the ions were required to be parallel or antiparallel to the field direction. Below T_c this parallel susceptibility falls off as shown in fig. 5. Above T_c eqn. (2.5) is again followed.

At the time neither the behaviour predicted by Néel nor that predicted by Bitter was known experimentally. However, Van Vleck (1941) extended these molecular field calculations and it was found that the theoretical predictions were now in encouraging agreement with the then recent experimental results on the compounds MnO, MnF_2, Cr_2O_3, CrSb, MnS, MnSe, MnTe. For this reason Van Vleck's formulation has come to be regarded as for many purposes the basic theory of antiferromagnetism. Therefore let us next briefly consider the more important equations in this theory.

2.2. *The Theory of Van Vleck*

We shall attempt an (approximate) calculation of the thermodynamic properties of a crystal model described by the Hamiltonian (2.4), assuming the presence of interactions between nearest neighbours only. Following Van Vleck we replace the true Hamiltonian (2.4) by the approximate form

$$V^{\mathrm{op}} = |J| \sum_i \mathbf{S}_i^{\mathrm{op}} \cdot \sum_{n=1}^{z} \langle \mathbf{S}_n^{\mathrm{op}} \rangle , \qquad \ldots\ldots(2.6)$$

where $\langle \mathbf{S}_n^{\mathrm{op}} \rangle$ is the, as yet undetermined, expectation value of the operator $\mathbf{S}_n^{\mathrm{op}}$ of one of the z nearest neighbours to atom i. This has the form of a sum of independent terms and so can be diagonalized immediately. Each atom i is now in an effective field (the 'molecular' field) $-2|J| \sum_{n=1}^{z} \langle \mathbf{S}_n^{\mathrm{op}} \rangle / g\beta$. Taking the direction of this field as the z direction we therefore know that $\langle S_{ix}^{\mathrm{op}} \rangle = \langle S_{iy}^{\mathrm{op}} \rangle = 0$ and that S_{iz}^{op} has eigenvalues $S, S-1, \ldots -S+1, -S$. But what is true for atom i is true for all the others, so that to the degree of approximation that (2.6) is the Hamiltonian, we can pick on an arbitrary direction in space and speak of the spin moment of any individual atom in this direction. This description, which ignores the non-commutation of S_{iz}^{op} with the true Hamiltonian (2.4) is sometimes called 'semi-classical' (*semi*-classical because S_{iz} is still allowed only discrete values). That the z-direction is arbitrary reflects the fact that the Hamiltonian (2.4) is completely isotropic.

We now come to a calculation of the average value of S_{iz} for a system in thermodynamic equilibrium. Van Vleck at this point restricts himself to crystal lattices for which none of the nearest neighbours of one atom are nearest neighbours of one another. This excludes among others the face-centred cubic and close packed hexagonal lattices. But the restriction allows us to divide the lattice into two interpenetrating sub-lattices (1 and 2) such that the nearest neighbours of sub-lattice 1 all lie on sub-lattice 2 and vice versa. In other words the exchange interactions of an atom on 1 are entirely with atoms on 2. Hence for a system in thermodynamic equilibrium the mean value of S_z for all atoms on sub-lattice 1 is the same (equal say to $\overline{S}_{1z}$) and likewise the average value of S_z is the same for all atoms on sub-lattice 2 ($\overline{S}_{2z}$). But from the fact that the molecular field on an atom of type 1 is

$$H' = -2|J| \Sigma \langle \mathbf{S}_n^{\mathrm{op}} \rangle / g\beta = -2|J| z \overline{S}_{2z} / g\beta , \qquad \ldots\ldots(2.7)$$

we can understand that the stable state is the one which actually makes this field positive, i.e. makes

$$\overline{S}_{1z} = -\overline{S}_{2z} \equiv \overline{S} , \qquad \ldots\ldots(2.8)$$

so that the average moments on the two sub-lattices are equal and opposite. The variation of $\bar{S}$ with temperature can be obtained by substituting the molecular field H' in place of H in formula (2.1). This gives the implicit equation

$$\bar{S} = S B_s\left(\frac{2|J|z S\bar{S}}{kT}\right). \qquad \ldots\ldots(2.9)$$

The solution $\bar{S}$ of this equation is S as $T \to 0$ but falls catastrophically to zero as T approaches a critical temperature T_c, given by

$$T_c = \tfrac{2}{3}|J|zk^{-1}S(S+1).$$

Above T_c the only solution is $\bar{S}=0$. This type of behaviour which is well known in the theory of ferromagnetism and other co-operative phenomena is shown

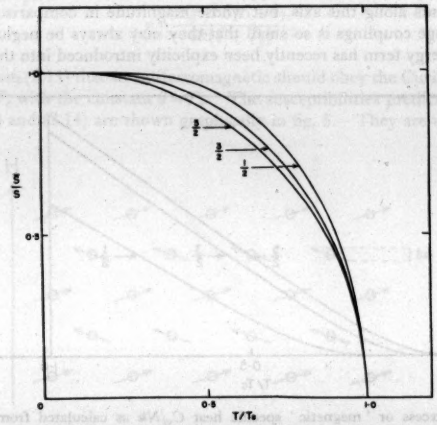


Fig. 2. Curves showing the variation of $\bar{S}/S$ with the reduced temperature T/T_c as calculated from eqn. (2.9). The numbers on the curves indicate the corresponding values of S. Notice that $\bar{S}$ is identically zero above T_c.

$$E = -N|J|z\bar{S}^2,$$

graphically in fig. 2. From (2.9) follows the magnetic internal energy and the magnetic or excess specific heat:

$$C_v = \frac{dE}{dT} = -2N|J|z\bar{S}\frac{d\bar{S}}{dT}. \qquad \ldots\ldots(2.10)$$

This is shown as a function of T/T_c in fig. 3. So far we have neglected any possible variations of J with temperature; but we would expect such variations to occur because the average separation between the atoms will increase as the temperature is raised. In this way the magnetic energy can make an important contribution to the coefficient of expansion. If the crystal expands isotropically then there is an additional term in the coefficient of expansion given by

$$\Delta\alpha = 2\kappa N z\bar{S}\frac{d\bar{S}}{dT}\frac{d|J|}{dv} \qquad \ldots\ldots(2.11)$$

where κ is the compressibility and $d|J|/dv$ is the derivative of the exchange integral with respect to the atomic volume. Since neither of these quantities would be expected to vary appreciably with temperature, we can infer that $\Delta\alpha$ will behave in the same way as the specific heat, namely that it will show a sharp peak at the Curie point.

Let us next consider the magnetic properties of this model. We notice that the result (2.8) only states that the moments of the two sub-lattices are antiparallel to one another and says nothing about their absolute direction in space. As we have already mentioned, this direction is strictly arbitrary if we use only the Hamiltonian (2.4). Nevertheless Van Vleck supposes there to be just one preferred axis along which the atomic moments are aligned. This is equivalent to postulating the presence of anisotropy forces which are just sufficient to keep the atomic moments along this axis, but whose magnitude in comparison with the strong exchange couplings is so small that they may always be neglected. (An anisotropy energy term has recently been explicitly introduced into the theory by

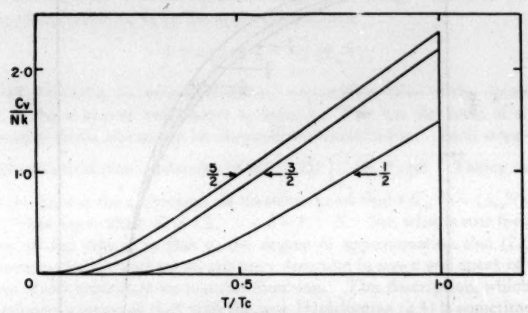


Fig. 3. The excess or 'magnetic' specific heat C_v/Nk as calculated from eqn. (2.1). Values of S are indicated.

Nagamiya 1951, and particularly clearly by Yosida 1951 a.) With this idea in mind we can legitimately speak of the orientation of an external magnetic field relative to the axis of spontaneous antiferromagnetism (z axis). The application of a field $\mathbf{H}$ will induce an equal moment in each sub-lattice; hence we can write

$$\bar{\mathbf{S}}_1 = \bar{\mathbf{S}} + \delta\mathbf{S}, \quad \bar{\mathbf{S}}_2 = -\bar{\mathbf{S}} + \delta\mathbf{S},$$

where we have used a vector notation for the mean moments because the axis of quantization for a given atom now depends on the sub-lattice it belongs to. The situation is represented schematically in fig. 4. The net moment of the whole crystal is then parallel to $\mathbf{H}$ and is given by $\mathbf{M} = \chi\mathbf{H} = Ng\beta\delta\mathbf{S}$. To determine $\delta\mathbf{S}$, i.e. $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$, Van Vleck applies (2.1) to each sub-lattice separately, replacing $\mathbf{H}$ by the appropriate effective field which is the vector sum of $\mathbf{H}$ and the molecular field due to the atoms on the other sub-lattice. He obtains the following results for the susceptibility in the limit of low fields:

(i) Below the Curie point χ depends on whether the applied field is parallel to the direction of spontaneous antiferromagnetism (case *a*) or whether it is perpendicular to this axis (case *b*).

Case a:
$$\chi_{\parallel} = \frac{Ng^2\beta^2S^2B_s'(y)}{k[T+3T_cS(S+1)^{-1}B_s'(y)]}, \qquad \ldots\ldots(2.12)$$

where $y = 2z|J|S\bar{S}/kT$.

Case b:
$$\chi_{\perp} = Ng^2\beta^2S(S+1)/6kT_c = Ng^2\beta^2/4z|J|. \qquad \ldots\ldots(2.13)$$

(ii) Above the Curie point there is no spontaneous antiferromagnetism and the crystal is isotropic.
$$\chi = \frac{Ng^2\beta^2S(S+1)}{3k(T+T_c)}. \qquad \ldots\ldots(2.14)$$

We notice from (2.14) that an antiferromagnetic should obey the Curie–Weiss law (2.5) above T_c with the constant $\theta = T_c$. The susceptibilities predicted by eqns. (2.12), (2.13) and (2.14) are shown graphically in fig. 5. They are unaltered by

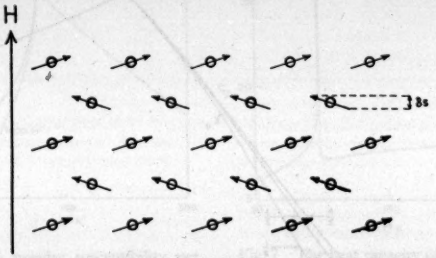


Fig. 4. A schematic representation of the effect of a magnetic field applied perpendicular to the axis of alignment of an antiferromagnetic. The axis of quantization for any particular atom is the direction of the resultant of **H** and the molecular field due to its neighbours.

the introduction of anisotropy terms except for a *slight* difference between $\chi_{\parallel}$ and $\chi_{\perp}$ *above* the Curie point, but this is insignificant compared with the difference, due to the exchange coupling, which exists below T_c (cf. Griffel and Stout 1950). Such a magnetic anisotropy is obviously easily detectable in single crystals. However, many antiferromagnetics are studied experimentally as powders, and these we must expect to contain a random distribution of antiferromagnetic axes and therefore to be isotropic. Averaging over all directions of the axis of spontaneous antiferromagnetism we find that the powder susceptibility is

$$\chi_p = \tfrac{2}{3}\chi_{\perp} + \tfrac{1}{3}\chi_{\parallel}. \qquad \ldots\ldots(2.15)$$

It follows therefore from (2.12), (2.13) and (2.15) that χ_p at $T=0$ is 2/3 of χ_p at $T=T_c$. This is verified approximately for a number of substances.

2.3. *Early Experimental Studies*

The first application of Van Vleck's theory was made not to the transition metals themselves—to which we originally appealed when introducing the idea of an antiferromagnetic—but to a number of their compounds. In this section we shall review the early experimental results which encouraged the belief that Van Vleck's model was a fair representation of these compounds. On the other hand the divergences between experiment and the simple theory were sufficiently obvious to stimulate developments not only of the mathematical treatment, but also of the basic model. The comparison between theory and experiment is complicated for some substances by a dependence of their properties on the mode of preparation; this leads to disagreement between the measurements of different investigators. For example, Haraldsen, Rosenqvist and Grønvold (1949) found values for the powder susceptibility of CrSb only one fourth of those reported earlier by G. Foëx and Graff (1939) for the same substance. This effect is probably

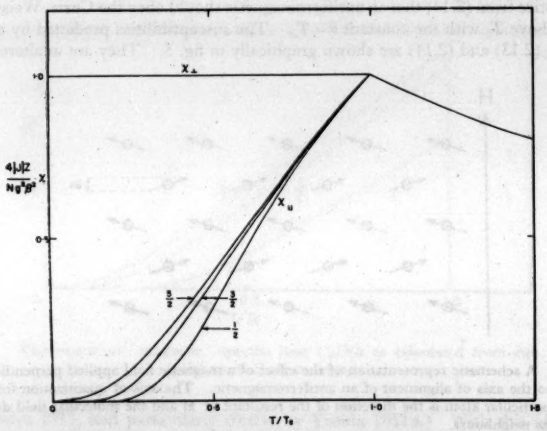


Fig. 5. Curves showing the dependence of the reduced susceptibility χ/χ_{T_c} on the reduced temperature T/T_c as predicted by eqns. (2.12), (2.13) and (2.14).

connected with the known ability of many transition metal compounds to depart from exact stoichiometric ratios (see Jost 1952, especially chap. II). A very striking example of the dependence of magnetic properties on composition is provided by ferrous sulphide; this is paramagnetic (possibly antiferromagnetic) at the stoichiometric composition but ferromagnetic when there is 10% excess sulphur (Yosida 1951 b).

Nevertheless Van Vleck's theory provided a general understanding of the susceptibility measurements which had been made by Bizette, Squire and Tsai (1938) on a powdered specimen of MnO. Their results are shown in fig. 6, and are seen to be of the same general form as the theoretical curves (fig. 5). We

observe that χ_0/χ_{T_c} is close to 2/3. Above T_c, χ_p was found to follow the Curie–Weiss law (2.5) but with a value of 610°K for the constant θ; this is five times as great as T_c although eqn. (2.14) calls for $\theta = T_c$. But it was in qualitative agreement with the theory that MnO showed a sharp peak in its specific heat (Millar 1928, see fig. 7); the value of T_c inferred this way—116°K—agrees well with that inferred for the magnetic measurements, namely 122°K.

Squire (1939) extended the magnetic measurements to the related compounds MnS, MnSe and MnTe, using for the last two some samples whose specific heat had already been determined by Kelley (1939). He found that the sulphide, like the oxide, has an angular maximum in the susceptibility; this occurs at 140°K, in exact agreement with the location of the specific heat maximum which had been found by C. T. Anderson (1931). More recent measurements of the susceptibility of MnS by Bizette and Tsai (reported by Bizette 1946), however, have yielded a Curie temperature somewhat higher, 165°K; this could conceivably be due to the use of a different specimen. Bizette and Tsai found that the Curie–Weiss law (2.5) was obeyed above T_c providing that θ was taken to be 528°K: this again is several times T_c.

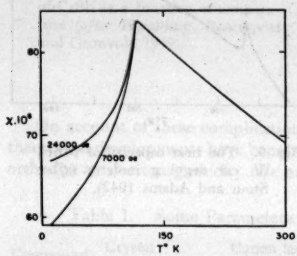


Fig. 6. The powder susceptibility per gramme of MnO as a function of the absolute temperature. We notice that, with the smaller field, χ_0/χ_{T_c} is very nearly 2/3 (*after* Bizette, Squire and Tsai 1938).

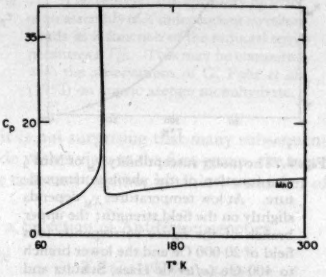


Fig. 7. The heat capacity of MnO (in cal deg^{-1} g mole^{-1}) (*after* Millar 1928).

Squire's results on MnSe at first sight also seemed to indicate substantial agreement with the Van Vleck theory, except for two things. Firstly, although he was using the same sample which Kelley (1939) had shown to have a pronounced specific heat maximum at 247°K, the susceptibility showed *its* maximum at 165°K. Secondly, Squire noticed an appreciable hysteresis in the dependence of the induced moment upon the applied field. Further studies of MnSe (Bizette and Tsai 1941 a, Lindsay 1951) have disclosed in addition a marked thermal hysteresis in the susceptibility. Lindsay proposed a very plausible explanation of his results, based on the assumption that there are really three forms of MnSe, stable in different temperature ranges. We shall not go into details, but we note that specimens of MnSe with zinc blende and wurtzite structures have been prepared and are unstable at room temperatures. The discrepancy between the thermal and the magnetic measurements is then explained if we suppose that the

anomalous specific heat indicates a phase transition and not an antiferromagnetic transition. There is an obvious need for detailed x-ray studies of MnSe, in order to test Lindsay's hypotheses.

The magnetic measurements on MnTe show that the susceptibility has an angular maximum at about 310°K—in complete agreement with Kelly's (1939) discovery of a maximum in the specific heat at 307°K. Kelley was unable to produce any supercooling below the transition temperature in MnTe, as indeed he was able to do for MnSe because of the thermal hysteresis.

Another manganese compound which had been investigated before Van Vleck's theory was published was MnF_2. Magnetic studies in substantial agreement with one another were made by Bizette and Tsai (1939) and by de Haas, Schultz and Koolhaas (1940), whose data are plotted in fig. 8. In fig. 9 we show the

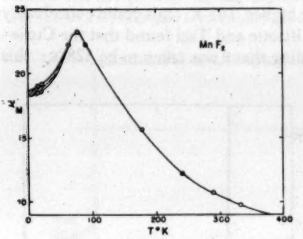


Fig. 8. The molar susceptibility χ_M of MnF_2 as a function of the absolute temperature. At low temperatures χ_M depends slightly on the field strength; the upper branch of the figure corresponds to a field of 20 000 Oe and the lower branch to 400 Oe (*after* de Haas, Schultz and Koolhaas 1940).

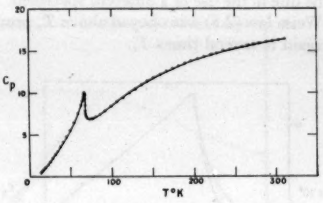


Fig. 9. The heat capacity of MnF_2 (in cal deg^{-1} g $mole^{-1}$) (*after* Stout and Adams 1942).

specific heat of MnF_2 as measured by Stout and Adams (1942). In MnF_2 the Mn ions are arranged on a body-centred tetragonal lattice, and it might therefore be expected to conform better to the Van Vleck theory than a substance like MnO where the magnetic ions lie on a face-centred cubic lattice—which cannot be divided into two sub-lattices. Actually χ_0/χ_{T_c} is about 0·8 and the parameter θ is 113°K compared with $T_c = 72$°K.

Without listing the studies of many more compounds it is probably already clear that some antiferromagnetics behave more simply and more like the predictions of the Van Vleck theory than do the others. Of the five manganese compounds just considered MnO, MnTe and MnF_2 stand out by virtue of the consistency and relative simplicity of the experimental data. Further examples are provided by FeF_2 (Bizette and Tsai 1941b), FeO (Bizette and Tsai 1943, Millar 1929) and CoO (Henry la Blanchetais 1951). On the other hand examples of substances whose behaviour is sufficiently different from theory to direct attention to their special features, are provided by MnSe, CrSb, V_2O_3 and VO_2. We show the susceptibility of CrSb in fig. 10 as measured by Haraldsen, Rosenqvist and Grønvold (1949); we note the low value of χ_0/χ_{T_c} (0·25) and that the Curie–

Weiss law is not obeyed above T_c. A similar behaviour was found in V_2O_3 by Hoschek and Klemm (1939) although here the rise in χ_p at the critical temperature (173°K) was even more abrupt. They also found this behaviour in VO_2 ($T_c = 341$°K) and noted that it was greatly affected by departures from the stoichiometric composition.

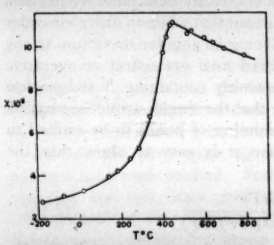


Fig. 10. The susceptibility per gramme of CrSb as a function of temperature (*after* Haraldsen, Rosenqvist and Grønvold 1949).

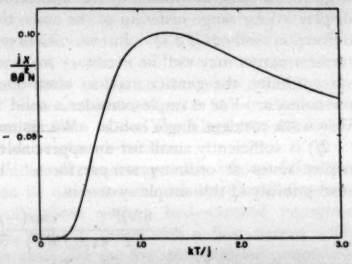


Fig. 11. The reduced susceptibility $j\chi/N\beta^2$ of an assembly of N independent covalent bonds as a function of the reduced temperatures kT/j. This may be compared with the observations of G. Foëx *et al.* (1953) on cupric acetate monohydrate.

On account of these complications it is not surprising that many subsequent theoretical developments have been made with a view to the first group, the more orthodox antiferromagnetics. We have collected together in table 1 a number of

Table 1. Some Parameters of a Selection of Antiferromagnetics

Compound	Crystal structure	Cation lattice structure	T_c(°K)	θ(°K)	χ_0/χ_{T_c}	g
MnF_2	Rutile	Body-centred Tetragonal	72	113	0·76	1·93
FeF_2	,,	,,	79	117	0·72	2·27
CoF_2	,,	,,	38	53	—	2·65
NiF_2	,,	,,	73	116	—	2·47
MnO_2	,,	,,	84	316	0·94	1·49
MnO	NaCl	Face-centred Cubic	122	610	0·67	2·01
MnS	,,	,,	165	528	0·82	1·98
FeO	,,	,,	198	570	0·79	2·88
CoO	,,	,,	292	280	—	2·03

Except for CoF_2 and NiF_2 the Curie point quoted here is that temperature at which the powder susceptibility is a maximum. CoF_2 and NiF_2 do not appear to be typical antiferromagnetics from the temperature variation of their powder susceptibility, but recent specific heat measurements (Stout and Catalano 1953), indicate that they have Curie points as quoted. In general, since the low temperature susceptibility depends slightly on the strength of the applied field, χ_0/χ_{T_c} is here referred to the weakest field used. The sources of the data are given in table 3.

the specifying parameters of the compounds in this group. This table shows values of T_c, θ, χ_0/χ_{T_c} and the gyromagnetic ratio g. The departures of g from 2 give a measure of the residual orbital moment. Crystal structures are also quoted;

further details can be obtained from the tables of Wyckoff. All the data refer to measurements on powder specimens.

The apparent lack of interest in the less orthodox antiferromagnetics is nevertheless unfortunate, because with only susceptibility (perhaps also thermal) data to go on, the sense in which they are 'antiferromagnetic' remains obscure. If they display a long range ordering of the spins this is directly detectable by neutron diffraction methods (§ 3.1). But we wish to point out that while an order–disorder transformation may well be necessary for producing an angular maximum in the susceptibility, the gentler maxima often observed *need* not reflect co-operative phenomena. For example consider a solid assembly containing N *independent* very weak covalent single bonds. We assume that the singlet–triplet separation ($\equiv 2j$) is sufficiently small for an appreciable number of bonds to be excited to triplet states at ordinary temperatures. Then it is easy to show that the susceptibility of this simple system is

$$\chi = \frac{8N\beta^2}{kT} \frac{\exp(-j/kT)}{\exp(j/kT) + 3\exp(-j/kT)}. \qquad \ldots\ldots(2.16)$$

As we see from fig. 11, χ is zero when T is zero and rises to a maximum at $T_c \simeq 5j/4k$. The Curie–Weiss law is not strictly obeyed until $T \gg T_c$: one then obtains $\theta \simeq 2T_c/5$. This situation appears to exist in cupric acetate monohydrate (G. Foëx, Karantassis and Perakis 1953). Although the detailed crystallography of this compound has not been determined by x-rays, the paramagnetic resonance spectrum corresponds to there being two paramagnetic Cu^{2+} ions in the unit cell (Bleaney and Bowers 1952). The exchange interaction, i.e. the 'weak covalent bond' may be assumed to be between these two ions. Another possible example is the solid known as Wurster's Blue perchlorate (tetra-methyl-p-phenylene-diaminium perchlorate: Elliott and Wolfsberg 1953). However, we note the fact that an assembly of independent systems can give a susceptibility curve which is not so very different from that accompanying the co-operative destruction of long-range order. This shows the need for caution in making inferences from susceptibility and specific heat data alone. Fortunately the technique of neutron diffraction, developed since the war, has provided very direct information about the magnetic state of a large number of substances. This forms part of the next section.

§ 3. Later Experimental Studies

3.1. *Neutron Diffraction**

We have seen how the theory of Van Vleck gave at least a qualitative understanding of the observed properties of a number of compounds. Conversely therefore we can say that the experimental results have indirectly confirmed the assumptions of the theory. However, the most satisfactory confirmation of the basic picture of an antiferromagnetic as an ordered alternating arrangement of spins has been provided only within the last five years by application of neutron diffraction methods (Shull and Smart 1949). When neutrons are incident upon a crystal lattice containing paramagnetic ions they are scattered by the atomic nuclei, as usual, and also by interaction of the neutron spin moment with the moment of the paramagnetic ions; this was shown by Halpern and Johnson (1939). An antiferromagnetically ordered arrangement of ionic moments, in other words a

* A general review of the applications of neutron diffraction was given in last year's *Reports on Progress in Physics* by Bacon and Lonsdale (1953).

magnetic super-lattice, will therefore give rise to neutron diffraction super-lattice lines in the same way as an ordered alloy gives x-ray diffraction super-lattice lines (Nix and Shockley 1938; antiferromagnetic ordering cannot of course be detected by x-rays since there is no interaction between the x-rays and the ionic moments). We would expect the magnetic super-lattice lines to diminish rapidly in intensity as the temperature is increased to near the Curie point, since at T_c all long range order is suddenly lost. These expectations have now been fully confirmed for a number of antiferromagnetics. The neutron diffraction patterns also give direct evidence as to the *type* of ordering present in the compound of interest; for example MnO, in which the Mn^{2+} ions are arranged on a face-centred cubic lattice and which therefore does not conform to the Van Vleck two-sub-lattice model, has been shown to be divided into eight sub-lattices.

We do not have space to consider all the compounds whose diffraction patterns have been studied. Instead let us review the results which have been obtained for two main groups: (i) those with a body-centred tetragonal arrangement of the magnetic ions, and (ii) those with a face-centred cubic arrangement. As the first group we shall take the five isomorphous compounds MnO_2, MnF_2, FeF_2, CoF_2* and NiF_2* which have been examined recently by Erickson (1952, 1953). All five compounds are antiferromagnetic and have the rutile type of crystal structure; the paramagnetic ions are therefore arranged on a body-centred tetragonal lattice (fig. 12). It is possible to divide this into two

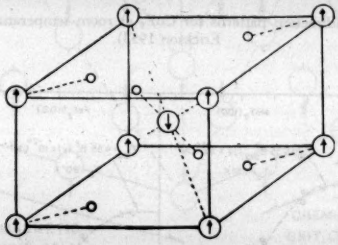


Fig. 12. A unit cell of the rutile lattice, with the c axis vertical. The large circles represent the cations and the small circles represent the anions. The arrows indicate the type of magnetic ordering found in MnF_2, FeF_2 and CoF_2. In NiF_2 the spin moments make a small angle of about 10° with the c axis, and in MnO_2 the ordering is of a more complicated type. VO_2 also has a rutile lattice but the magnetic structure is unknown (*after* Erickson 1953).

sub-lattices, one containing the body centre ions and the other the corner ions. However these substances do not quite conform to the Van Vleck model in that the *nearest* neighbours to a given cation are not the eight corresponding corner cations, but two cations one on each side along the (short) c axis. Nevertheless

* On the basis of powder susceptibility determinations extending down to 20°K, Bizette (1946) has stated that these two substances are not antiferromagnetic. The variation of $1/\chi_p$ with T appears to be almost linear above about 30°K with no sign of the characteristic antiferromagnetic minimum. Nevertheless the neutron diffraction results and some experiments on single crystals to be described in §3.2, make it quite plain that both NiF_2 and CoF_2 do, in fact, show antiferromagnetic order–disorder transitions even though the powder susceptibility may be somewhat unorthodox.

although the (8) corner ions are some 20% farther away than the (2) nearest neighbours it appears that the next nearest neighbour interactions are of greater importance (except MnO_2). In fact Erickson's data for MnF_2, FeF_2, CoF_2, and NiF_2 are accounted for with a magnetic unit cell of the same dimension as the chemical unit cell, the moments of the magnetic ions at the lattice corners being antiparallel to the moments of the body-centred ions. Typical diffraction patterns (actually for CoF_2) are shown in fig. 13; they show the presence of a

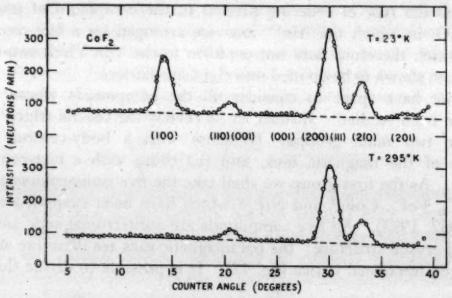


Fig. 13. Neutron diffraction patterns for CoF_2 at room temperature and at 23°K (*after* Erickson 1953).

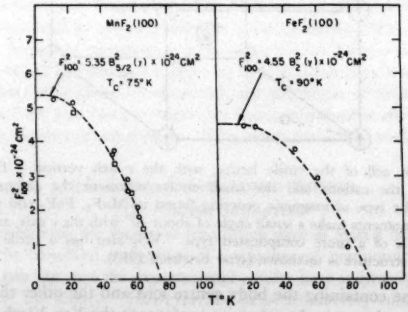


Fig. 14. Temperature dependence of the (100) peak intensity, in terms of the crystal structure factor for MnF_2 and FeF_2. The broken lines represent the solutions of eqn. (2.9) for $S=5/2$ and 2 respectively (*after* Erickson 1953).

super-lattice line at 23°K but not at 295°K, i.e. well above the Curie temperature for CoF_2. For all four compounds MnF_2, FeF_2, CoF_2, and NiF_2 Erickson found that the intensity of this main super-lattice line varied with the temperature almost exactly as Van Vleck's theory would predict for ions with S respectively equal to 5/2, 2, 3/2 and 1. The experimental data for MnF_2 and FeF_2 are

compared with theory in fig. 14. The agreement is perhaps rather surprising in view of the simplified assumptions underlying the theory. On the basis of this agreement however Erickson has inferred Curie temperatures and gyromagnetic ratios for the four fluorides. They are respectively MnF_2, 75°K, 1·99; FeF_2, 90°K, 2·32; CoF_2, 50°K, 2·00; NiF_2, 83°K, 2·04. There is some uncertainty in the g-value quoted for NiF_2, since the spin moments of the Ni^{2+} ions appear to make a small angle with the c axis, unlike the Mn^{2+}, Fe^{2+} and Co^{2+} moments which are aligned along the c axis. The g-values may be compared with those inferred by applying eqn. (2.5) to the high-temperature powder susceptibility data shown in table 1.

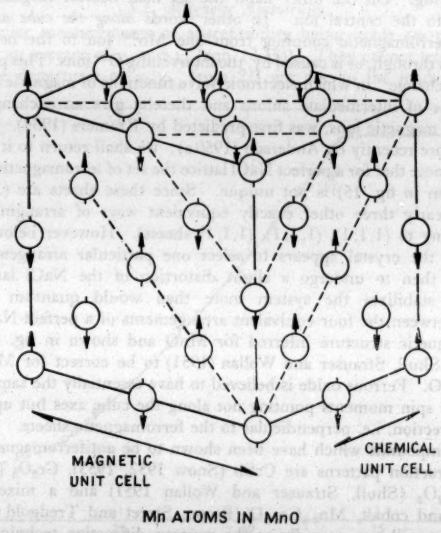


Fig. 15. Antiferromagnetic structure existing in MnO below its Curie temperature of 120°K (only the Mn^{2+} ions are marked). The arrangement of spins shown here is one of four equivalent but distinguishable arrangements; each is specified by a set of parallel (111) planes, shown by the broken lines in the diagram. There are obviously four such sets of planes (*after* Shull, Strauser and Wollan 1951).

Erickson's work demonstrates the existence of substances conforming to the two-sub-lattice model, but earlier studies had already disclosed the existence of more complicated types of antiferromagnetic ordering. Shull, Strauser and Wollan (1951) examined MnO and a number of other compounds (MnS, MnSe, FeO, CoO, NiO) crystallizing in the NaCl type of structure.* In such a structure the cations are located on a face-centred cubic lattice and thus would not be expected to conform to the two-sub-lattice theory. Nevertheless magnetic

* See §3.3 for qualification of this statement.

super-lattice lines were found in all six compounds. Shull, Strauser and Wollan considered in detail a number of possible magnetic structures for MnO but only the one shown in fig. 15 was compatible with the observed diffraction patterns. The structure is seen to consist of parallel sheets (111 planes) within which all the Mn^{2+} ions are arranged ferromagnetically, but with the antiferromagnetic coupling between neighbouring sheets. Of the twelve nearest neighbours surrounding any Mn^{2+} ion one half is oriented parallel to the central ion and the other half is oriented antiparallel; with this arrangement any nearest neighbour interactions will give the same contribution to the energy as would a completely random distribution, so that such interactions cannot be the source of the ordering. On the other hand the six next nearest neighbours are all antiparallel to the central ion. In other words *along the cube axes* there is always antiferromagnetic coupling from one Mn^{2+} ion to the next and this coupling acts through, or is caused by, the intervening O^{2-} ions. This phenomenon of 'super-exchange', in which electronic wave functions of magnetic ions (Mn^{2+}) overlap those of intermediate anions and thereby give an exchange coupling between the magnetic ions, was first predicted by Kramers (1934). It has been discussed more recently by Anderson (1950a); we shall return to it in § 4.3.

We may note that for a perfect NaCl lattice the set of ferromagnetically aligned sheets (shown in fig. 15) is not unique. Since these sheets are (1$\bar{1}$1) planes, there are clearly three other exactly equivalent ways of arranging the spins (corresponding to (1, $\bar{1}$, 1), ($\bar{1}$, 1, 1), ($\bar{1}$, $\bar{1}$, 1) sheets). However below the Curie temperature the crystal appears to select one particular arrangement (set of sheets) and then to undergo a slight distortion of the NaCl lattice. This presumably stabilizes the system more than would quantum mechanical resonance between the four equivalent arrangements of a perfect NaCl lattice.*

The magnetic structure inferred for MnO and shown in fig. 15 was also inferred by Shull, Strauser and Wollan (1951) to be correct for MnS, MnSe, CoO and NiO. Ferrous oxide is believed to have essentially the same structure, but with the spin moments pointing not along the cube axes but up and down the (111) direction, i.e. perpendicular to the ferromagnetic sheets.

Other compounds which have been shown to be antiferromagnetic by their neutron diffraction patterns are CrSb (Snow 1952, 1953), Cr_2O_3 (Brockhouse 1953), α-Fe_2O_3 (Shull, Strauser and Wollan 1951) and a mixed oxide of manganese and cobalt, $Mn_{0.5}Co_{0.5}O$ (Bacon, Street and Tredgold 1953). No doubt the list will grow rapidly as the neutron diffraction technique becomes more widely applied. It is of particular interest to study substances whose status still remains in doubt after measurement of magnetic and thermal properties. For example the susceptibility of CuO passes through a *diffuse* maximum between 100°C and 200°C (Klemm and Schüth 1931); we have already referred to this kind of departure from the Van Vleck theory at the end of § 2. Then again Hoschek and Klemm (1939) found that the susceptibility of VO_2 varied very little with temperature below 68°C but at this point it sharply increased

* Actually the Hamiltonian (2.4) has no matrix elements between the wave functions corresponding to these four arrangements of spin. However, the correct wave function should also contain terms corresponding to all intermediate configurations of the spins; these would link the four principal terms together. But considerations due to P. W. Anderson (1951), given in § 4.4, indicate that the intermediate arrangements would be present to only a small degree and that therefore resonance is unlikely to lower the energy very much.

to seven times its former value and then with further rise in temperature decreased very slowly; the (χ, T) curve, in fact, was almost a step function. Neutron diffraction experiments can tell us whether this is yet another variety of antiferromagnetic order–disorder transition. It would also be of considerable interest to examine compounds such as $FeCl_2$, $CoCl_2$, $NiCl_2$ which appear to be normal antiferromagnetics except for a pronounced dependence of the susceptibility and the specific heat anomaly on the applied field (Bizette 1946, Shalyt 1939 a, b). They are also characterized by having negative values of θ (Starr, Bitter and Kaufmann 1940); these may be explicable in terms of secondary ferromagnetic interactions (§ 4.2).

3.2. *Magnetic Anisotropy*

The earlier measurements of susceptibility being made on powders only determined an average of $\chi_{\parallel}$ and $\chi_{\perp}$ (eqn. (2.15)). They cannot tell us therefore whether the departures of χ_0/χ_{T_c} from 2/3 are due to the inadequacy of the

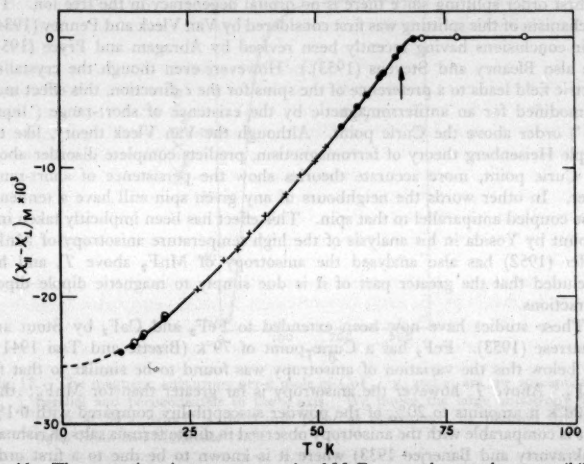


Fig. 16. The magnetic anisotropy per g mole of MnF_2. $\chi_{\parallel}$ and $\chi_{\perp}$ are the susceptibilities parallel and perpendicular to the c axis. The arrow indicates the temperature of the maximum in the heat capacity (*after* Griffel and Stout 1950).

postulate of a single antiferromagnetic axis or to some other weakness in the theory. In view of the way that the fluorides MnF_2, FeF_2, CoF_2, NiF_2 seem to fit the two sub-lattice model (§ 3.1), we shall now consider the independent measurement of $\chi_{\parallel}$ and $\chi_{\perp}$, made by Stout and others. Actually they measured only the anisotropy $\chi_{\parallel}-\chi_{\perp}$, but taken with a knowledge of the powder susceptibility both $\chi_{\parallel}$ and $\chi_{\perp}$ can easily be found separately. The anisotropy was determined by measuring the torque necessary to balance that exerted on a single crystal by the homogeneous field of an electromagnet. All four compounds are tetragonal and were therefore mounted with the c axis perpendicular

to the axis of rotation (since there is no anisotropy in a plane perpendicular to the c axis).

The Van Vleck theory tells us that the anisotropy $\chi_{\parallel}-\chi_{\perp}$ is almost entirely due to the way the exchange forces lock the spins parallel and antiparallel to a preferred crystal direction. We can see from fig. 5 that $\chi_{\parallel}-\chi_{\perp}$ is zero at and above the Curie point but must become increasingly large and negative as the temperature falls from T_c to zero. The behaviour of MnF_2 as observed by Griffel and Stout (1950) is shown in fig. 16. Since the susceptibility is least along the c axis it follows that this is the direction in which the spins are aligned —in complete agreement with the neutron diffraction results. The anisotropy of MnF_2 is not quite zero above the Curie point but is positive and of the order of 0·1% of the susceptibility. This decreases slowly with temperature and is similar to the anisotropy observed in dilute manganous salts (Krishnan and Banerjee 1936) a second order effect which is due to a slight splitting of the ^6S level of the free Mn^{2+} ion by the crystalline electric field. (There is of course no first order splitting since there is no orbital degeneracy in the free ion. The mechanism of this splitting was first considered by Van Vleck and Penney (1934); their conclusions having recently been revised by Abragam and Pryce (1951). See also Bleaney and Stevens (1953).) However, even though the crystalline electric field leads to a preference of the spins for the c direction, this effect must be modified for an antiferromagnetic by the existence of short-range ('liquid like') order above the Curie point. Although the Van Vleck theory, like the simple Heisenberg theory of ferromagnetism, predicts complete disorder above the Curie point, more accurate theories show the persistence of short-range order. In other words the neighbours of any given spin still have a tendency to be coupled antiparallel to that spin. This effect has been implicitly taken into account by Yosida in his analysis of the high temperature anisotropy of MnF_2. Keffer (1952) has also analysed the anisotropy of MnF_2 above T_c and has concluded that the greater part of it is due simply to magnetic dipole–dipole interactions.

These studies have now been extended to FeF_2 and CoF_2 by Stout and Matarrese (1953). FeF_2 has a Curie point of 79°K (Bizette and Tsai 1941 b) and below this the variation of anisotropy was found to be similar to that for MnF_2. Above T_c however the anisotropy is far greater than for MnF_2; thus at 300°K it amounts to 20% of the powder susceptibility compared with 0·1%. This is comparable with the anisotropy observed in dilute ferrous salts (Krishnan, Chakravorty and Banerjee 1933) where it is known to be due to a first order interaction between the orbital motion of the electrons and the crystalline electric fields. This in turn indicates that the orbital momentum of the Fe^{2+} ion ($(3d)^6$, ^5D) is not completely quenched when it enters the crystal FeF_2, in agreement with the large g-value which we inferred from the powder susceptibility (table 1). On this basis one might expect an even larger high temperature anisotropy in CoF_2 since the free Co^{2+} ion is in an F-state. In fact the results given by Stout and Matarrese (1953) indicate that the magnetic behaviour of CoF_2 is determined as much by the residual orbital momentum as by the exchange interactions (which are reduced anyway because S is only 3/2 for Co^{2+}). We can see this from fig. 17. Although $\chi_{\parallel}-\chi_{\perp}$ is least at 0°K and rises smoothly to a nearly vanishing value above 150°K, only 60% of the anisotropy has vanished at the Curie point, which is 50°K according to the neutron diffraction results.

If the anisotropy were due entirely to the exchange interactions it would have vanished 100% by the Curie point.

Although considerable attention has been given to the way that crystalline electric fields and spin–orbit coupling affect the energy levels and magnetic behaviour of ions in *dilute* paramagnetic salts (see for example Van Vleck 1940 or Bleaney and Stevens 1953) few attempts have been made to apply this work to cases in which exchange interactions between the ions are also important. Yosida (1952) and also Gorter and Haantjes (1952) have however made this application for the special case of the antiferromagnetic $CuCl_2 . 2H_2O$ ($T_c = 4{\cdot}3°K$). The crystalline electric field removes the fivefold orbital degeneracy of the Cu^{2+} ion (^2D) and it is assumed that the resultant splitting is sufficiently great for the population of levels other than the ground level to be negligible. This simplifies the problem and Gorter and Haantjes then proceed by replacing the '*g*-factor'

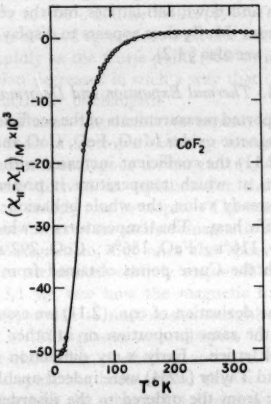


Fig. 17. The magnetic anisotropy per g mole of CoF_2. $\chi_{\parallel}$ and $\chi_{\perp}$ are the susceptibilities parallel and perpendicular to the *c* axis. We notice the large anisotropy which exists at 50°K, which is the Curie point as determined by neutron diffraction (*after* Stout and Matarrese 1953).

with a *g*-tensor (Pryce 1950) and using a molecular field method to treat the exchange interactions. Yosida examines a slightly more detailed model and includes the dipole–dipole interactions between the spins. In this way both authors are able to explain the critical field effects discovered by N. J. Poulis *et al.* (1951). However, the simplification that all the ions are in the lowest orbital level at the temperature of interest does not appear to hold for CoF_2. It is well known that paramagnetic salts containing Co^{2+} ions do not obey Curie's law (Fowler and Guggenheim 1939, §1414). This is believed to indicate an appreciable excitation of Co^{2+} ions from the ground level to higher levels of the orbital multiplet produced by the crystalline electric field. The models used in the theory of antiferromagnetism do not appear to have been extended to include this possibility yet. As indicated at the beginning of §2 they mostly

assume complete 'quenching' of the orbital momentum, which as we may see from Pryce's paper (1950) means a very strong crystalline field producing large separations between the terms of the orbital multiplet (Λ_{ij} negligibly small).

We conclude this section by noting that in addition to MnF_2, FeF_2 and CoF_2 the anisotropy of the isomorphous MnO_2 has also been studied (Bizette and Tsai 1949) with interesting results. Unlike the fluorides where the spin moments are aligned along the tetragonal c axis, in MnO_2 they appear to be arranged perpendicular to it thus giving a temperature independent susceptibility along the c axis below T_c. However although the susceptibility in a plane perpendicular to the c axis falls off quickly as the temperature is lowered beyond T_c the crystal is isotropic in this plane. These results and also the neutron diffraction results (Erickson 1952) are explained by assuming that the crystal is divided into four magnetic sub-lattices, two in which the spins are up and down and two in which the spins are left and right. Referring to fig. 12 we may say that corner ions in MnO_2 belong to up and down sub-lattices but the centre ions belong to the left and right sub-lattices. MnO_2 thus appears to display 'ordering of the second kind' (Van Vleck 1951, see also §4.2).

3.3. *Thermal Expansion and Deformation*

In 1948 M. Foëx reported measurements of the coefficient of thermal expansion of the four antiferromagnetic oxides MnO, FeO, CoO and NiO. As one expects on the basis of eqn. (2.11) the coefficient increases with increasing temperature below the Curie point, at which temperature it presents a sharp maximum. Above T_c it falls to a steady value, the whole behaviour being qualitatively the same as that of the specific heat. The temperatures at which the maxima occurred were as follows: MnO, 116°K; FeO, 186°K; CoO, 292°K; NiO, 523°K. These may be compared with the Curie points obtained from the susceptibility data (table 1).

We recall that in the derivation of eqn. (2.11) we assumed that all the lattice parameters change in the same proportion or in other words that there is no *distortion* of the crystal lattice. Early x-ray diffraction studies of MnO, MnS and FeO by Ellefson and Taylor (1934) were indeed unable to detect any changes in lattice type in going from the ordered to the disordered condition. Recent studies have, however, provided conclusive evidence that slight distortions do take place when (some) antiferromagnetics are cooled below their Curie points.

Following earlier work by Rooksby (1948) which had demonstrated the existence in NiO of a small rhombohedral deformation from the face-centred cubic structure below about 200°C, Tombs and Rooksby (1950) undertook the study of MnO, FeO and CoO. The first two were shown to be slightly rhombohedral at temperatures below their Curie points (face-centred cubic above) but in CoO the distortion was found to be tetragonal. The measurements on CoO were repeated by Greenwald and Smart (1950), who were able to demonstrate accurately the equivalence of the Curie temperature and the temperature at which the lattice begins to be tetragonal. Recently Rooksby and Tombs (1951) have added MnS to the list of those face-centred cubic antiferromagnetics which are known to show a rhombohedral distortion at low temperatures. An interesting study by Willis and Rooksby (1953) has shown that the magnitude of the rhombohedral distortion in FeO depends on the deficiency of iron. The change by 90°K was found to be almost negligible for

oxides whose iron content was less than 48 atomic % (the rhombohedral angle was within several minutes of the 60° which corresponds to an undistorted cubic lattice). Closely analogous changes have been observed in some non-cubic antiferromagnetic compounds, in particular those with hexagonal and rhombohedral lattices. In such lattices distortion occurs simply by an alteration in the axial ratio without the modification of symmetry. Greenwald (1951) has shown that Cr_2O_3, which has a rhombohedral lattice, undergoes a marked contraction in the [111] direction when cooled below the Curie point ($\simeq 45°$c), while the isomorphous α-Fe_2O_3 undergoes an expansion along the [111] direction at its Curie point 675°c (Willis and Rooksby 1952). The difference in behaviour between the two compounds is the same as that existing between FeO, which takes a rhombohedral angle slightly less than 60°, and MnO and NiO where the angle becomes slightly greater than 60°. The lattice parameters of CrSb, which has the hexagonal NiAs structure, have been studied by Snow (1953) throughout the temperature range 0°c to 800°c and by Willis (1953) from $-200°$c to 600°c. The two investigations are in good agreement with one another and show that the c-dimension increases rapidly as the Curie point (450°c) is approached. At the same time the a-dimension decreases in such a way that the cell volume, rather surprisingly, remains practically unchanged.

At first sight it would seem impossible to explain the deformations from cubic symmetry on the basis of isotropic exchange interactions. Nevertheless, as we shall see, the situation is analogous to that existing with the susceptibility; we recall that Van Vleck derived an equation for the magnetic anisotropy, $\chi_{\parallel} - \chi_{\perp}$, of a model which strictly was isotropic. The reason he was able to do this is that the anisotropic forces in an antiferromagnetic determine the direction of spontaneous antiferromagnetism, but once the spins have picked out this axis, the exchange forces are almost entirely responsible for the susceptibility anisotropy. Now in §3.1 we saw how the magnetic ordering in MnO (also MnS, MnSe, CoO, NiO) is due to next nearest neighbour interactions and is such that there is a set of (111) planes with all spins in a given plane parallel, though the spins in neighbouring planes are antiparallel (fig. 15). Once the particular alternation of spins has been settled this set is unique. The three other sets of planes defined by the other [111] directions correspond to distinct but completely equivalent alternations of spins. However we must suppose that the existence of small strains and imperfections in the lattice will lead a crystal to prefer one set of planes to another. Smart and Greenwald (1951) have pointed out that expansion or contraction of the lattice in the direction perpendicular to the sheets, i.e. in the resultant direction of the antiferromagnetic interactions, will deform the lattice and give it rhombohedral symmetry. Such an expansion (or contraction) is furthermore to be expected since the exchange integral J will depend on distance and the magnetic energy will therefore give an expansion (or contraction) depending on temperature, very much as one would expect for a perfectly isotropic dilatation, eqn. (2.11). However, the fact that FeO expands while MnO and NiO both contract must therefore imply a difference in sign between the corresponding dJ/dx, x being the separation between planes. Such a difference is unintelligible on the simple Heisenberg model where dJ/dx is always of the same sign for antiferromagnetics, although it may be of either sign for ferromagnetics (see for example Seitz 1940, p. 614, fig. 12). The super-exchange mechanism results in a more complicated form

for J (eqn. (4.10)) which may well account for the different behaviour of FeO and NiO, but so far the variations of J with distance have not been discussed.

An interesting corollary to all this is the possible formation of 'domains'. It seems hardly likely that the imperfections in a single crystal will be such as to favour completely one set of (111) sheets. Rather we should expect the crystal to divide into regions each of which is specified by one of the four equivalent sets of (111) sheets. The corresponding distortions of the lattice will make this division very real, and we may reasonably expect it to give rise to a decrease in the moduli of elasticity and also an increase in the internal friction. Both effects have recently been observed in CoO† (Street and Lewis 1951, Fine 1953).

3.4. *Antiferromagnetic Resonance*

The techniques of magnetic resonance absorption which have yielded detailed and precise information on the paramagnetic salts (Bleaney and Stevens 1953) have not so far been very illuminating when applied to antiferromagnetics. This is due to the occurrence of very high resonance frequencies lying at the limit of the experimentally accessible range. The situation is easily understood on the basis of Kittel's (1951) calculation for the Van Vleck two-sub-lattice model. If H_A be a suitable internal field representing the anisotropy forces, and if γ be the magneto-mechanical ratio $g\epsilon/2mc$, then Kittel shows that an antiferromagnetic below its Curie point should have two natural frequencies ω_0 given by

$$\omega_0/\gamma = H \pm [H_A(H_A + 2H')]^{1/2}, \qquad \ldots\ldots(3.1)$$

where, as in § 2, H and H' are the external and molecular fields respectively. For materials with moderately high Curie points such as MnF_2, H' is of the order of 10^6 Oe and is much greater than either H or H_A. Hence (3.1) reduces to a single frequency

$$\omega_0 = \gamma(2H_A H')^{1/2}.$$

Now Keffer (1952) finds that the anisotropy field in MnF_2 at 0°K is almost 9000 Oe, and hence $\omega_0 \simeq 10$ cm^{-1}. This is a much larger frequency than the 0·3 cm^{-1} used by Maxwell and McGuire (1953) and explains why their (paramagnetic) spin resonance absorption lines become extinguished as the compound is cooled below its Curie point. This phenomenon is observed in Cr_2O_3, MnO, MnS, MnSe and MnTe. The g-values inferred from the paramagnetic resonances are 1·97 for Cr^{3+} and 2·00 in all four manganese compounds. It is interesting to note that the extinction of the spin resonance in MnSe showed a thermal hysteresis similar to that exhibited by the susceptibility (Lindsay 1951).

Ubbink *et al.* (1952) have, however, been able to observe the antiferromagnetic resonance doublet in $CuCl_2.2H_2O$, because this compound has a very low Curie point (4·3°K, Van den Handel *et al.* 1952) and therefore a comparatively low molecular field. $CuCl_2.2H_2O$ has rhombic symmetry so that the resonance frequencies, as well as static properties such as the susceptibility, are functions of the orientation of the applied steady field relative to the crystal axes. The theory of antiferromagnetic resonance in a rhombic crystal has been successfully worked out by Ubbink (1953), who applied the static molecular field theory

† The neutron diffraction experiments have shown that CoO has the same magnetic structure as MnO, yet it undergoes a tetragonal deformation below T_c, not rhombohedral as do MnO, MnS, FeO and NiO. Thus CoO can have only three and not four types of domain.

due to Gorter and Haantjes (1952). Accounts of this work have been given recently by Gorter (1953 a, b).

§4. Theoretical

4.1. *Introduction*

The approximate theory which we gave in § 2.2 was designed for the simplest possible model, one with only two sub-lattices and with interactions only between nearest neighbours. Such a model was found to obey the Curie–Weiss law above the Curie temperature providing that θ is put equal to T_c. We have seen though that there are very few substances for which this equality is even roughly satisfied. Now Néel (1936 b) in the original application of the molecular field approximation to antiferromagnetism showed how the existence of next nearest neighbour (n.n.n.) interactions would lead to values of θ different from T_c (see also Néel 1948). More recent studies of the effect of n.n.n. interactions have been made by P. W. Anderson (1950 b) and by Smart (1952), paying particular

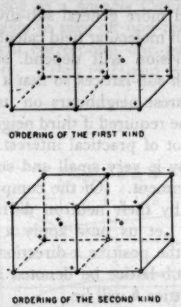


Fig. 18. The two types of antiferromagnetic order possible in a b.c.c. lattice with only n.n. and n.n.n. interactions. If both interactions are antiferromagnetic in sign, the upper arrangement is stable when $J(\text{n.n.})/J(\text{n.n.n.})>3/2$; otherwise, the lower arrangement is stable (*after* Smart 1952).

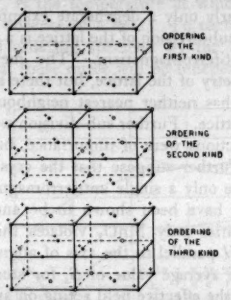


Fig. 19. The three types of antiferromagnetic order possible in a f.c.c. lattice with only n.n. and n.n.n. interactions. If both interactions are antiferromagnetic then only the second and third types are stable. Order of the third kind occurs if $J(\text{n.n.})/J(\text{n.n.n.})>2/3$, otherwise it is of the second kind (as observed in MnO). Order of the first kind only occurs if the n.n.n. interactions are ferromagnetic (*after* Smart 1952).

attention to body-centred and face-centred cubic (f.c.c.) lattices. It is obvious that if the n.n.n. interactions become sufficiently large then they, and not the nearest neighbour (n.n.) interactions, determine the type of order which sets in below the Curie point. Anderson has made a detailed study of this effect and has shown for example that in f.c.c. lattices θ/T_c must always lie between 5 and 1 if only antiferromagnetic couplings are present. These calculations have been reviewed previously by Van Vleck (1951), Street (1951) and Smart (1953 a), so we shall present only a rather condensed account of them here (§ 4.2). In § 4.3 we consider the mechanism of super-exchange, which accounts for the

large n.n.n. interactions responsible for the type of order observed in MnO and related compounds (cf. §3.1). Despite the great value of Anderson's work in enabling us to understand the observed θ/T_c ratios, his calculations are still based on the molecular field approximation which replaces the Hamiltonian (2.4) by (2.6). The more accurate descriptions of a model specified by the Hamiltonian (2.4) form the material of §4.4.

4.2. *Extended Molecular Field Calculations*

In considering the extension of the Van Vleck model by inclusion of n.n.n. interactions let us begin with the simple case of a body-centred cubic (b.c.c.) lattice. If the n.n.n. interactions are small, the type of ordering will be unaltered and will be as shown in the upper half of fig. 18. On the other hand, if these interactions are large and antiferromagnetic they will oppose this arrangement and the ground state will be as shown in the lower half of fig. 18; this requires four, and not two, sub-lattices for its description. Analytically, however, it is possible to handle both cases together by using four sub-lattices, since the first is clearly only a degenerate example of the second more general sub-division. This sub-division of the lattice is a preliminary to all molecular field calculations in antiferromagnetism. The method of sub-division will depend on the symmetry of the lattice, but there should be enough sub-lattices so that a given atom has neither nearest neighbours nor next nearest neighbours on its own sub-lattice. Further sub-division would obviously be required if third neighbour interactions were of importance, but this case is not of practical interest. We shall further suppose that the crystalline anisotropy is very small and such as to give only a single antiferromagnetic axis of alignment. (Of the compounds which have been shown to be antiferromagnetic by their neutron diffraction intensities only MnO_2 violates this restriction.) Let us now apply a small field H parallel to the axis of alignment (defining the positive z-direction) and let the average value of S_z for atoms on the lth sub-lattice be denoted by $\bar{S}_l$. Then the effective field acting on an atom of sub-lattice k will be

$$H_k = H + \sum_{l \neq k} \Gamma_{kl} \bar{S}_l, \qquad \ldots\ldots(4.1)$$

in the z-direction. Here $\Gamma_{kl}\bar{S}_l$ is the effective field acting on an atom of the kth sub-lattice due to its neighbours on the lth sub-lattice. The coefficients Γ_{kl} are proportional to the corresponding exchange integrals (cf. eqn. (2.7)) and may easily be written down for any given lattice, but we shall not require them explicitly here. By inserting (4.1) into (2.1) we obtain a set of equations from which to obtain the $\bar{S}_k$ and hence also the susceptibility $\chi_{\parallel}$.

$$\frac{\bar{S}_k}{S} = B_s\left[\frac{g\beta S}{kT}\left(H + \sum_{l \neq k} \Gamma_{kl}\bar{S}_l\right)\right]. \qquad \ldots\ldots(4.2)$$

When $\bar{S}_k/S \ll 1$, i.e. either above the Curie temperature or in the immediate vicinity below it, we can expand the Brillouin function in (4.2) and retain only the first term. This gives

$$\bar{S}_k = \frac{g\beta S(S+1)}{3kT}\left[H + \sum_{l \neq k} \Gamma_{kl}\bar{S}_l\right]. \qquad \ldots\ldots(4.3)$$

Let f_k be the fraction of (magnetic) atoms on the kth sub-lattice, then the susceptibility above T_c is easily seen to be

$$\chi = \frac{Ng\beta}{H}\sum_k f_k \bar{S}_k = \frac{Ng^2\beta^2 S(S+1)}{3k(T+\theta)}, \qquad \ldots\ldots(4.4)$$

where

$$\theta = \frac{g\beta S(S+1)}{3k}\sum_{l\neq k} -\Gamma_{kl}. \qquad \ldots\ldots(4.5)$$

Notice that these last two equations only follow if $\Sigma_{l\neq k}\Gamma_{kl}$ is independent of the index k, something which is obviously true if all the sub-lattices are equivalent. If not all the sub-lattices are equivalent, as in the ferrites, then the Curie–Weiss law (4.4) will not be obeyed (cf. Fairweather, Roberts and Welch 1952, especially §4). If now we set H equal to zero in (4.3), non-zero solutions $\bar{S}_k$ are possible only if the determinant of the coefficients vanishes, i.e.

$$\det\left|\delta_{kl} - \frac{g\beta S(S+1)}{3kT}\Gamma_{kl}(1-\delta_{kl})\right| = 0. \qquad \ldots\ldots(4.6)$$

This equation determines a number of temperatures, each of which must be considered as a *possible* Curie point, since T_c is the temperature at which $\bar{S}_k$ just begins to take on non-zero values in the absence of a field. To each root there will correspond one particular kind of ordering specified by the ratios $\bar{S}_k/\bar{S}_l$ etc. From (4.2) and (4.3) it is clear that these ratios will be independent

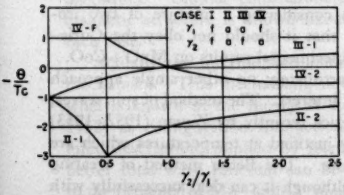


Fig. 20. θ/T_c values for a b.c.c. lattice as calculated from eqns. (4.5) and (4.6). The roman numerals on each curve refer to the signs of the interactions (f, ferromagnetic, a, antiferromagnetic); the arabic numerals refer to the kind of magnetic ordering (fig. 18). γ_1 and γ_2 are molecular field coefficients proportional to the exchange integrals and such that $\gamma_2/\gamma_1 = 3J(\text{n.n.n.})/4J(\text{n.n.})$ (*after* Smart 1952).

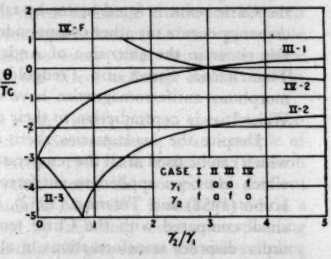


Fig. 21. θ/T_c values for a f.c.c. lattice as calculated from eqns. (4.5) and (4.6). The roman numerals on each curve refer to the signs of the interactions (f, ferromagnetic, a, antiferromagnetic); the arabic numerals refer to the kind of magnetic ordering (fig. 19). γ_1 and γ_2 are molecular field coefficients proportional to the exchange integrals and such that $\gamma_2/\gamma_1 = J(\text{n.n.n.})/2J(\text{n.n.})$ (*after* Smart 1952).

of temperature, i.e. the average moment on sub-lattice k divided by its value at $T=0$ varies with temperature exactly as shown for the Van Vleck model (fig. 2). This means that the thermal properties of the two models are also identical. The real Curie temperature may be taken as the largest root of (4.6) since this corresponds to the arrangement of lowest energy at absolute zero (Yafet, private communication). However, this rule assumes the impossibility

of transitions from one type of order to another at some intermediate temperature. As demonstrated by Smart (1953 a) such transitions are indeed impossible if the molecular field coefficients Γ_{kl} are independent of temperature. In practice, although there is little empirical evidence for such transitions, the thermal expansion of the lattice can conceivably alter the Γ_{kl} sufficiently to give rise to them (Smart 1953 b).

For particular lattices, such as the b.c.c. and the f.c.c., symmetry allows the solutions of (4.6) to be obtained easily. In fig. 19 we show the three types of antiferromagnetic order corresponding to the three distinct antiferromagnetic solutions for the f.c.c. lattice. The order observed in MnO (fig. 15) is an example of the second kind. In figs. 20 and 21 we show the ratio θ/T_c as a function of the relative strengths of the n.n. and n.n.n. interactions for b.c.c. and f.c.c. lattices. Comparison with table 1 shows that the observed θ/T_c ratios do in fact all lie within the range allowed theoretically.

Although these molecular field calculations have provided an illuminating correlation between the neutron diffraction results and the susceptibility they are not quantitatively very satisfactory. Thus Howard and Smart (1953) have shown that they still lead to that *broad* peak in the specific heat which the simple Van Vleck theory gives (fig. 3) and which we have seen to be in poor agreement with the *sharp* peaks observed in practice (cf. fig. 7). Furthermore although the high temperature susceptibility of MnO requires $\theta/T_c = 5$, the susceptibility below the Curie point is fitted better by taking $\theta/T_c = 1$ (Smart 1952). A similar discrepancy occurs for other compounds such as MnS, FeO, and CoO. Difficulties also occur in the extension of Anderson's theory to mixed antiferromagnetics. Thus Bacon, Street and Tredgold (1953) considering a mixture of two isomorphous antiferromagnetics have shown that it should not obey the Curie–Weiss law, in contradiction to their own experimental results on MnO + CoO.

Despite the inadequacies there is unfortunately no other single approach which can be used at all the temperatures of interest. The method of spin waves, which has been applied to antiferromagnetics recently by Ziman (1952, 1953) Kubo (1952) and Tessman (1952), is only justified at temperatures which are small compared with the Curie temperature. The Bethe method of treating order–disorder transformations in alloys, although it can deal successfully with the Ising ferromagnet or antiferromagnet, unfortunately breaks down at low temperatures when applied to the Heisenberg–Van Vleck model* (P. W. Anderson 1950 c, Li 1951). It ought not to be surprising therefore if the crude molecular field method continues to be used in the theoretical study of more complicated antiferromagnetics. One fairly obvious way to extend the models that we have been using is to do away with the assumption that all the magnetic

* The Ising (1925) model replaces the Hamiltonian (2.4) by $2J\Sigma_{i,j} S_{i_\alpha}{}^{op} S_{j_\alpha}{}^{op}$ so that $S_{i_\alpha}{}^{op}$ is a constant of the motion. It is generally studied under the additional restriction that S is $\frac{1}{2}$, and the statistical problem then becomes formally identical with that for the formation of a super-lattice in an AB-type alloy. *In the molecular field approximation* there is no essential difference between the Ising and the Heisenberg models of a ferromagnetic, but this equivalence does not hold for an *anti*-ferromagnetic. This follows from the fact that the Ising Hamiltonian selects a single z-direction which is the same for all atoms whereas (2.4) does not. The distinction is unimportant when all the spins are parallel, but a glance at fig. 4 shows its significance when they are antiparallel. *Exact* solutions of the Ising problem have been obtained for a number of two-dimensional lattices (Wannier 1950, Newell and Montroll 1953).

ions are effectively in S-states and to introduce the effects of the orbital momentum by using the results of the theory of paramagnetic salts (Van Vleck 1940, Bleaney and Stevens 1953). We have already indicated the need for such an extension in connection with CoF_2. The simpler case of $CuCl_2.2H_2O$ has been studied by Gorter and Haantjes (1952, see also Gorter 1953 a) and by Yosida (1952), using molecular field methods. This work has been applied to the problem of antiferromagnetic resonance absorption by Ubbink (1953) and again by Yosida (1952).

4.3. *Super-exchange*

We have seen empirically how the important spin aligning forces in many antiferromagnetics (e.g. MnO and MnS) appear to act from one cation to another *through* an intermediate anion. This in itself is an indication that the interaction is not a simple exchange interaction—a conclusion which may also be reached on general theoretical grounds. For, as is well known, such a direct exchange interaction is supposed to depend in magnitude and sign on the extent of the overlap of the 3d electron wave functions. The interaction favours antiparallel spins only when the (magnetic) ions are nearer together than a critical distance; at greater separations the interaction is ferromagnetic, but its strength falls off rapidly at large separations (Seitz 1940, especially pp. 613–4). Now in FeO, for example, the n.n.n. Fe^{2+} ions are 4·28 Å apart whereas in ferromagnetic α-iron the separation is only 2·86 Å. Therefore on the conventional Heitler–London–Heisenberg picture it is inconceivable that direct exchange interaction between the Fe^{2+} ions in FeO should be either antiferromagnetic in sign or of any appreciable magnitude. Similar considerations applied to every other compound whose magnetic structure is known likewise rule out direct exchange interactions in each case.

Actually the intermediate anion plays a very active part in the interaction mechanism. This mechanism, now called 'super-exchange', is an example of what is known in the theory of atoms (Hartree 1947) and molecules (Coulson 1952) as configuration interaction. In other words the completely divalent ionic description, implied by formulae such as $Mn^{2+}O^{2-}$, is inadequate and a better total wave function can be obtained by including terms corresponding to the presence of Mn^{+} and O^{-} ions. That this will give rise to the spin coupling we require was shown first by Kramers and Bloch (Kramers 1934).

Following Kramers let us consider a simple model in which the magnetic ions have only one electron spin each (e.g. Cu^{2+}) and in which the anions have completely filled shells (e.g. O^{2-}). Let us suppose also that the crystalline electric field has completely removed all *orbital* degeneracy in the ions, so that in the absence of a magnetic field each cation is just doubly degenerate (due to its spin). Lastly we shall assume that one-electron wave functions (orbitals) centred about different nuclei are orthogonal. Then if N be the number of cations in the whole solid and if there are no (direct) exchange interactions present the ground configuration will be 2^N-fold degenerate, since we can assign each cation either an α spin function (spin up) or a β spin function (spin down). Corresponding to each assignment (t) of all N spins we can form a determinantal wave function Ψ_t using the N* one-electron product functions (space × spin).

* In addition we must include the one-electron anion wave functions although, of course, they will be the same for all the 2^N degenerate states belonging to the ground level. In what follows it is assumed that the anion functions have been included.

The 2^N different Ψ_t will differ only in the assignment of spin functions, and any one of them is an equally good wave function for the solid. In order to understand more clearly how configuration interaction gives the super-exchange coupling, let us first digress and consider what happens if some small direct exchange interaction is introduced as a perturbation. As usual we proceed by taking an arbitrary linear combination of the Ψ_t,

$$\Phi = \Sigma_t a_t \Psi_t$$

and solving the secular problem for the coefficients,

$$\Sigma_{t'} H_{tt'} a_{t'} = E a_t, \qquad t = 1, 2, \ldots\ 2^N.$$

E is the energy and $H_{tt'}$ is the matrix element of the total Hamiltonian H^{op} between Ψ_t and $\Psi_{t'}$. It will be found that all the non-diagonal elements $H_{tt'}$ are zero, except those for which the spin assignment t' differs from t only by the interchange of the (opposite) spins on two orbitals. This way of approaching the exchange perturbation problem (due to Slater 1929) is completely equivalent to the Dirac spin operator method (Corson 1951, especially chap. 11) with which we started in § 2. In fact the matrix $H_{tt'}$ is—apart from a constant diagonal term—identical with the matrix of the operator

$$V^{\text{op}} = 2J \Sigma_{i,j} \mathbf{S}_i^{\text{op}} \cdot \mathbf{S}_j^{\text{op}}, \qquad \ldots\ldots (4.7)$$

with respect to a certain set of 2^N mutually orthogonal spin functions. Each of these functions corresponds to a particular determinant Ψ_t, and is obtained from it by multiplying together the N *one-electron* spin functions which appear on the diagonal of Ψ_t.

However, let us now return to the original problem in which there occur no direct exchange interactions but in which we have agreed to mix the wave functions of the ground configuration with those of the next highest configuration. The lowest excited configuration may be obtained by removing an electron from a particular anion and then placing it on a neighbouring cation. Such a configuration will be enormously degenerate; in addition to the spin degeneracy, there is now a choice of $z_c N$ ways in which to locate the excitation, z_c being the number of nearest neighbouring cations to a given anion. Let us denote the degenerate excited determinantal functions by Ψ_u. Then we can write the trial wave function as

$$\Phi = \Sigma_t a_t \Psi_t + \Sigma_u b_u \Psi_u,$$

and obtain the secular equations

$$\Sigma_{t'} H_{tt'} a_{t'} + \Sigma_u H_{tu} b_u = E a_t, \qquad \ldots\ldots (4.8\,a)$$

$$\Sigma_t H_{ut} a_t + \Sigma_{u'} H_{uu'} b_{u'} = E b_u. \qquad \ldots\ldots (4.8\,b)$$

Now if direct exchange interactions are negligible ($J = 0$) all the matrix elements $H_{tt'}$ will be zero except possibly the diagonal terms. As these are all equal to one another we may re-define the zero of energy so that they too are zero. Kramers solves the equations (4.8) by successive approximations and obtains as the (third order) equation for the a_t

$$\sum_{t'} \left[-\sum_u \frac{H_{tu} H_{ut'}}{\Delta E} + \sum_{\substack{u,u' \\ u \neq u'}} \frac{H_{tu} H_{uu'} H_{u't'}}{(\Delta E)^2} \right] a_{t'} = E a_t,$$

where $\Delta E = H_{uu}$ is the excitation energy from the ground to the excited level. Detailed consideration of the terms in the square brackets now shows that the first is of the form $F \delta_{tt'}$, and hence may be absorbed by a re-definition of

the scale of E, while the second term represents a matrix of the same kind as $H_{tt'}$, and therefore corresponds to a coupling of type (4.7) between the spins. That configuration interaction leads to a ground state splitting which can be described as a coupling between the magnetic moments is not an isolated result, but one which may be related to the general extension of the Dirac vector model by inclusion of configuration interaction (P. W. Anderson 1950a). Nevertheless the role of the intermediate states can be seen very clearly in our simple model. Consider three atoms of the crystal A, B, C, such that A and C are magnetic (one outer electron each) and B is a non-magnetic neighbour to both A and C. In the ground state A and C will each be doubly degenerate; t could be written as $A^{1/2}B^0C^{-1/2}$. A state u can then be obtained by taking one of the electrons on B and pairing it with the electron already on A; $u \equiv A^0B^{1/2}C^{-1/2}$. Direct exchange of the antiparallel spins on B and C ($H_{uu'}$) gives the state $u' \equiv A^0B^{-1/2}C^{1/2}$. The final state t' is obtained now by taking an electron from A and pairing it with the one on B; $t' \equiv A^{-1/2}B^0C^{1/2}$. The net effect of going from t to t' is thus to interchange a (single) pair of initially antiparallel spins; this, as we have already stated above, is the condition that the element $H_{tt'}$ is not zero. Continuing in this way we can see the complete equivalence of

$$\sum_{\substack{u,u' \\ u \neq u'}} \frac{H_{tu}H_{uu'}H_{u't'}}{(\Delta E)^2}$$

to a matrix of the kind $H_{tt'}$, i.e. to the matrix of an operator of the type (4.7).

Such considerations, although they show us the origin of super-exchange interactions do not tell us under what conditions we would expect them to be antiferromagnetic. Nor do they immediately tell us why, in MnO for example, the interaction between two Mn ions is stronger when the bonds to the intervening O ion are collinear (n.n.n.) than when they are at right angles (n.n.), although from this fact one might infer that the super-exchange is principally due to the 2p electrons of the O ion and therefore possesses the same directional properties as the 2p function. A detailed consideration of the required extensions of the simple Kramers model has been given by P. W. Anderson (1950a). He allows the magnetic ions to have spins greater than 1/2, and lets the excited electron from the anion enter the neighbouring cation with its spin either parallel or antiparallel to the original magnetic moment. The corresponding energies of excitation will in general be different from one another; let us call them E_p and E_a respectively. (In Kramers' model E_p is effectively infinite, so that the added electron is always antiparallel.) Then Anderson shows that the super-exchange interaction gives a coupling between the spins of each pair of (cation) neighbours to a given anion which is represented by the operator

$$V_{ij}^{\text{op}} = 2J\mathbf{S}_i^{\text{op}} \cdot \mathbf{S}_j^{\text{op}}, \qquad \ldots\ldots(4.9)$$

where

$$J = \rho^2 K\left[\frac{1}{E_a^2} - \frac{1}{E_p^2}\right]. \qquad \ldots\ldots(4.10)$$

ρ is the 'resonance integral' for the excitation of an electron from the orbital ϕ on the anion to orbital ψ on the first cation, i.e.

$$\rho = \int \psi h^{\text{op}} \phi \, d\tau,$$

where h^{op} is an appropriate one-electron (Hartree) Hamiltonian for this electron. K is the exchange integral between the remaining electron in orbital ϕ and the

magnetic electrons on the second cation. If ϕ is a p-type orbital then either K or ρ will be small unless the two cations are on opposite ends of a diameter through the anion. In compounds with the NaCl structure (e.g. MnO, MnS etc.) this means that the principal interactions will be between n.n.n. cations—as observed. However, the corresponding s-orbital could also be excited, though with slightly more difficulty, to give an isotropic coupling. This very probably contributes to the n.n. interactions in the NaCl-type antiferromagnetics and is of importance in compounds not having cubic lattices (e.g. MnF_2).

It still remains to discuss the sign of J. Consider the factor $(E_a^{-2}-E_p^{-2})$. For an ion such as Cr^{2+} where the 3d shell is less than half full, Hund's rule tells us that $E_a>E_p$ or that $(E_a^{-2}-E_p^{-2})<0$. For ions with more than half-filled shells the additional electron can only enter the 3d shell if it has an antiparallel spin. If it has parallel spin then it must go into the 4s orbital which, it is asserted, lies at a higher energy than 3d,* so that $E_a<E_p$ and $E_a^{-2}-E_p^{-2}>0$. Hence Anderson concludes that the relative sign of J depends on whether the cation 3d shell is less than or more than half filled. It has been suggested (Van Vleck 1951, Street 1951) that this conclusion is amply confirmed by experiment. It is pointed out that CrTe is ferromagnetic whereas MnTe is antiferromagnetic (both have the NiAs lattice). It is also stated that VCl_2 and $CrCl_2$ are ferromagnetic but that $MnCl_2$ and $FeCl_2$ are antiferromagnetic. This however is incorrect; the measurements of Starr, Bitter and Kaufmann (1940) show that $CrCl_2$ is almost certainly antiferromagnetic ($T_c\simeq 40°K$) and that although VCl_2 may not be antiferromagnetic, its susceptibility does not depend on the field strength as one would expect it to if it were ferromagnetic. As additional evidence in favour of Anderson's conclusion one could add that CrO_2 is ferromagnetic (Michel, Chaudron and Bénard 1951) that MnO_2 is antiferromagnetic and that both have the rutile lattice. However the isomorphous VO_2 is probably antiferromagnetic whereas it should be ferromagnetic, like CrO_2, on Anderson's scheme. One may also draw attention to the oxides V_2O_3, Cr_2O_3 and α-Fe_2O_3 which have the rhombohedral lattice; these are *all* antiferromagnetic. We must therefore conclude that the qualitative inferences which have been made from (4.10) about the sign of J cannot be considered as justified. This is not of course a criticism of the super-exchange mechanism itself, although Slater (1953) has suggested that neglect of the non-orthogonality of neighbouring orbitals (e.g. ϕ and ψ) may be a serious objection. The applicability of Anderson's results (4.9) and (4.10) has also been questioned by Polder (1951) on the grounds that all the five 3d orbitals have been treated as equivalent. It is stated that if one takes account of their differences and also investigates the effect of the crystalline electric field in determining the distribution of electrons among the orbitals, then antiferromagnetic interactions are possible for ions with fewer than five electrons.

* This point is made rather implicitly and is not discussed. It is not true for the *free* ions Mn^+ and Fe^+, as may be seen from the tables of Bacher and Goudsmit (1932). They tell us that the ground configurations of these ions are respectively $3d^5\,4s$ and $3d^6\,4s$ and that the spin of the 4s electron is strongly coupled parallel to the moment of the 3d shell. Hence for the free iron and manganese ions $E_p<E_a$. On the other hand the ground configurations of the Co^+ and Ni^+ ions are $3d^8$ and $3d^9$ respectively, so that here $E_a<E_p$. But in the solid the large 4s orbit will be compressed by the neighbouring anions and it may be argued that this will reverse the sequence of levels in iron and manganese thus making $E_a<E_p$ here also. No calculations have been made however.

4.4. *More Accurate Calculations*

In this section we shall consider some of the improvements which have been made over the molecular field approximation. These studies, for the most part, have been made using a two sub-lattice model with interactions only between ions on different sub-lattices and such a model is to be assumed in what follows. There are many aspects to the problem of improving on the molecular field approximation. One may, for example, be interested in knowing how far the expectation value of S_z^{op} for an atom on the positive sub-lattice differs from S, supposing the whole system to be in its ground state (at absolute zero). A simpler problem, which we shall consider first, is to find how the true ground state energy E_0 differs from the energy given by the molecular field approximation, namely,

$$E_{\text{MF}} = -Nz|J|S^2. \qquad \ldots\ldots(4.11)$$

This expression may be obtained either by substituting S for $\bar{S}$ in the Van Vleck equation for the internal energy or by taking the diagonal matrix element of (2.4) with respect to the spin wave function which corresponds to the perfectly alternating arrangement. By the energy variation principle we can therefore say that (4.11) provides an upper bound to E_0. P. W. Anderson (1951) has pointed out that we can also find a *lower* bound which turns out to be

$$-Nz|J|S^2\left(1+\frac{1}{zS}\right). \qquad \ldots\ldots(4.12)$$

In many practical cases the range of variation allowed by (4.11) and (4.12) is very narrow; thus for MnF_2, $z=8$ and $S=5/2$, so that the possible fractional variation for E_0 is only one twentieth. This is a good indication that the ground state wave function is not greatly different from that which corresponds to the alternating arrangement with $S_{iz}=\pm S$ according as the atom i is on sub-lattice 1 or sub-lattice 2.

The reasoning by which Anderson arrives at (4.12) is fairly general so let us digress for a moment to consider it. The Hamiltonian (2.4) may be decomposed as follows,

$$V^{\text{op}} = \Sigma_i V_i^{\text{op}},$$

where

$$V_i^{\text{op}} = |J|\Sigma_j \mathbf{S}_i^{\text{op}} \cdot \mathbf{S}_j^{\text{op}}.$$

The summation runs over all the atoms j which are nearest neighbours to atom i. Suppose now that we know the representation which makes V^{op} diagonal and that we form the matrices $\mathcal{V}_i$ of the V_i^{op}. Then $\mathcal{V}=\Sigma_i\mathcal{V}_i$. The matrix $\mathcal{V}$ of V^{op} is diagonal but the $\mathcal{V}_i$ are not. Now the general variation theorem for Hermitian matrices tells us that even the lowest diagonal element of $\mathcal{V}_i$ is greater than the least eigenvalue of $\mathcal{V}_i$, which we know to be $-|J|S(zS+1)$ from the work of P. R. Weiss (1948). In symbols,

$$\mathcal{V}_i^{ll} > -|J|S(zS+1), \qquad \text{all } l.$$

Therefore

$$\Sigma_i\mathcal{V}_i^{ll} \equiv \mathcal{V}^{ll} > -N|J|S(zS+1), \qquad \text{all } l.$$

But $\mathcal{V}^{ll}$ is one of the eigenvalues of V^{op}. Hence all the eigenvalues of V^{op} are greater than $-N|J|S(zS+1)$ and thus

$$-Nz|J|S^2\left(1+\frac{1}{zS}\right) < E_0 < -Nz|J|S^2. \qquad \ldots\ldots(4.13)$$

That the ground state of an antiferromagnetic can be approximately described by a wave function which divides the crystal into two sub-lattices, one specified by $S_{iz}=+S$ and the other by $S_{jz}=-S$, is the guiding principle behind recent

spin wave theories of antiferromagnetism at low temperatures (P. W. Anderson 1952, Ziman 1952, 1953, Kubo 1952, Tessman 1952; for an interesting interpretation of these more mathematical papers see Keffer, Kaplan and Yafet 1953). The method of spin waves was originally developed by Bloch (1930, 1932) for the theory of ferromagnetism and was reformulated in a very elegant manner in 1940 by Holstein and Primakoff (H.P.). The general H.P. approach is the same for ferromagnetism and antiferromagnetism. Starting from that basic ordered state of the lattice which we get by making the molecular field approximation, we define spin deviation operators (e.g. $S - S_{iz}^{\text{op}}$) to describe the disordered states and then express the Hamiltonian H^{op} in terms of them. The resulting expression may then be expanded as a series in the deviation operators. If we now drop all terms higher than those of the first order in the spin deviations we obtain an approximate Hamiltonian, H_1^{op}; this should be good for describing the *lowest states* of the system since these states will have only small spin deviations. In general it is now possible, by a Fourier transformation on the deviation operators, to separate H_1^{op} into a set of independent harmonic oscillator terms whose eigenvalues can be written down immediately. Each oscillator corresponds to a wave of spin deviations which is propagated through the crystal by the exchange coupling between neighbours. Therefore in the ground state of the system the quantum number of each oscillator is zero. The corresponding ground state energy is not in general the same as the energy of that basic ordered state which guides our considerations towards the approximate Hamiltonian H_1^{op}. For a simple Heisenberg ferromagnetic with no anisotropy forces, it does happen that the two are equal to one another and to the exact ground-state energy. However, if we include magnetic dipole interactions between the spins the two energies are no longer equal and neither is exact. The difference between them may be interpreted in terms of zero-point motion of the spin waves. In antiferromagnetics also there is an appreciable effect due to the zero point oscillations. Thus Kubo (1952) and Ziman (1952) using the H.P. technique have independently calculated the ground levels of several simple lattices. Their results for the simple cubic (s.c.) and body-centred cubic lattices may be written as follows:*

$$E_0(\text{s.c.}) = -Nz|J|S^2\left(1 + \frac{0{\cdot}582}{zS}\right),$$

and

$$E_0(\text{b.c.c.}) = -Nz|J|S^2\left(1 + \frac{0{\cdot}584}{zS}\right),$$

with $z = 6$ and 8 respectively. It is encouraging that these expressions lie almost midway between the limits allowed by (4.13). The corresponding expectation values of S_{iz}^{op} are given by

$$\langle S_{iz}^{\text{op}}\rangle(\text{s.c.}) = \pm(S - 0{\cdot}078)$$

and

$$\langle S_{iz}^{\text{op}}\rangle(\text{b.c.c.}) = \pm(S - 0{\cdot}075),$$

the sign depending on whether atom i is on the positive sub-lattice or the negative sub-lattice. The ground levels of several antiferromagnetic lattices had also been calculated slightly earlier by P. W. Anderson (1952), using a simpler but less exact formulation of the spin wave approach; his results are identical with

* These formulae are valid in the limit of vanishingly small anisotropy forces. In the spin wave theories, as in the molecular field theories, some anisotropy is necessary in order to select the antiferromagnetic axis, but apart from this it *may* not be important.

those obtained by the H.P. technique. The two formulations do not however lead to identical results at non-zero temperatures.* Thus Kubo (1952) finds that the low temperature susceptibility parallel to the antiferromagnetic axis is

$$\chi_{\parallel} = \frac{f N g^2 \beta^2 \mathbf{k}^2 T^2}{z^3 |J|^3 S^2}, \qquad \ldots\ldots (4.14)$$

where the numerical factor f is $\sqrt{3}/8$ for a s.c. lattice and 1/6 for a b.c.c. lattice. On the other hand, although Tessman (1952) using Anderson's method also finds a T^2 dependence, the numerical coefficient is different: for a simple cubic lattice it is always less than in (4.14). Differences also exist over $\chi_{\perp}$. To the same degree of approximation that $\chi_{\parallel}$ is given by (4.14), Kubo finds that $\chi_{\perp}$ is independent of temperature and given by

$$\chi_{\perp} = N g^2 \beta^2 / 4z|J|. \qquad \ldots\ldots (4.15)$$

This is identical with the molecular field result (2.13). On the other hand Hulthén (1936), using essentially Anderson's formulation, found

$$\chi_{\perp} = \frac{N g^2 \beta^2}{24|J|}\left\{1 - \frac{1}{3\sqrt{3}}\left(\frac{\mathbf{k}T}{2S|J|}\right)^2\right\},$$

for a simple cubic lattice. The perpendicular susceptibility of MnF_2 is in fact observed to decrease with increasing temperature but this is probably due to other effects. On the basis of either formulation it follows that the magnetic anisotropy $\chi_{\parallel} - \chi_{\perp}$ at low temperatures is a linear function of T^2, a relation which is observed to hold for MnF_2 (Griffel and Stout 1950). Let us assume that (4.14) and (4.15) are the correct equations to use and then infer $|J|$ and hence T_c. If we use Van Vleck's formula for T_c we arrive at a theoretical Curie point of 62°K. This is in satisfactory agreement with the observed Curie point (72°K from the susceptibility maximum or 66·5°K from the specific heat maximum). A linear relation between $\chi_{\parallel} - \chi_{\perp}$ and T^2 is not found in FeF_2 or in CoF_2; empirically T^4 seems to be more appropriate (Stout and Matarrese 1953).

The formulae (4.14) and (4.15) are valid only in the limit of vanishingly small anisotropy forces. Tessman has shown however that when the anisotropy terms become sufficiently important $\chi_{\parallel}$ tends to zero much more rapidly than T^2. The decay is in fact of a similar exponential form to that given by the Van Vleck theory (fig. 5).

The fundamental approximation of replacing the true Hamiltonian H^{op} by the first order spin wave Hamiltonian H_1^{op} unfortunately restricts spin wave theories to temperatures which are 'small' compared with the Curie temperature. Quantitatively the limitations are not well understood, although in principle they can be investigated by a consideration of higher order Hamiltonians. It is certain however that the spin wave theories are not applicable in the region around the Curie point where all long range order has been lost. Practically the only attempt at an accurate high temperature theory is that of Li (1951) who has employed the Bethe–Weiss method (P. R. Weiss 1948). In this method the interactions in a cluster containing an arbitrary atom (the central atom) and all its immediate neighbours (the z 'first shell' atoms) are treated exactly and the

* In Anderson's formulation there occur quantities Q_λ, P_λ which are taken to be canonical conjugates. In fact Q_λ, P_λ are only approximately conjugate unless the spin deviations are all zero. The H.P. formulation does not include any such uncertainty, which is therefore probably the reason for the different predictions at non-zero temperatures.

exchange energy of this group is diagonalized. The interactions between the first shell atoms and their outside neighbours are then replaced by an internal field, the analogue of the molecular field in the Van Vleck theory. In the two sub-lattice model we shall, of course, have two types of cluster depending on which sub-lattice the central ion belongs to. Correspondingly we shall have two internal fields, which Li considers to be parallel or antiparallel to the applied field. The average magnetic moment of the central atom and that of the first shell atom can now be derived from the partition function of the cluster. Self-consistency requires that the magnetic moment of a central atom on a given sub-lattice be equal to that of a first shell atom on the same sub-lattice. In this way we obtain two relations from which to determine the two internal fields.

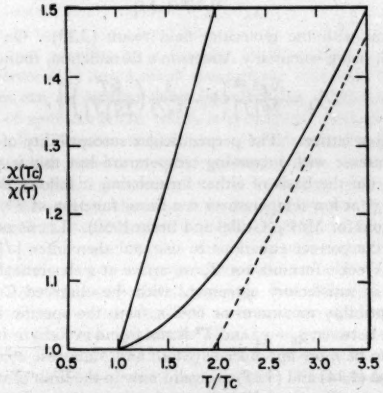


Fig. 22. The inverse of the susceptibility as a function of temperature for the case of a simple cubic lattice and $S=\frac{1}{2}$. The straight line was obtained on the basis of Van Vleck's theory (eqn. (2.14)) and the curve was calculated by Li using the Bethe–Weiss method. The broken line is the high temperature asymptote to Li's curve. Owing to the relative scale of temperature this figure does not show the difference between the corresponding Curie points (*after* Li 1951).

Li has used the above method to examine the antiferromagnetic behaviour of a number of simple lattices assuming that each atom has spin one half. He finds that antiferromagnetic order–disorder transitions will not occur in the linear chain, the honeycomb net, the quadratic layer or in the hexagonal layer. By qualitative arguments based on the results for the quadratic layer, he also infers that the face-centred cubic lattice (with only n.n. interactions) will not support antiferromagnetism—a result which has since been confirmed by Ziman (1953) on the basis of spin wave theory. This result does not contradict the fact that antiferromagnetism is observed in compounds with NaCl structure (e.g. MnO) for in these cases the primary interactions are between next nearest neighbours. It is also of interest that the values of $kT_c/|J|$ for the s.c. and b.c.c. lattices are substantially smaller than the values given by the molecular

field method; they are respectively 2·004 and 3·18 compared with 3 and 4. Although the molecular field method is probably more accurate than this when S is greater than one half, it suggests that the agreement between the observed Curie point in MnF_2 and the value calculated from the low temperature anisotropy should be treated with reserve.

Of greatest interest perhaps is the temperature variation of the susceptibility above the Curie point. In fig. 22 we show χ_{T_c}/χ plotted against T/T_c as calculated by Li for the simple cubic lattice. The Curie–Weiss law is obeyed asymptotically as $T \to \infty$, with θ_{lim} as given by the molecular field theory (this is an exact result, Opechowski 1937, 1938, Zehler 1950). θ_{lim} is not, of course, any longer equal to T_c. At temperatures closer to the Curie point the inverse susceptibility curve may be fitted reasonably well by a straight line over limited ranges of temperature. The value of θ obtained in this way will depend on the particular range which is selected, being largest when the range lies close to T_c and least ($\equiv \theta_{\text{lim}}$) when the temperatures are very high. By fitting a straight line in the region between $2T_c$ and $3T_c$, Li finds that θ/T_c should be 1·7 for the simple cubic lattice and 1·5 for the b.c.c. lattice; the corresponding limits θ_{lim}/T_c are 1·50 and 1·26. It is interesting to compare the value 1·5 with the observed ratios 1·57 for MnF_2 and 1·48 for FeF_2 (table 1) and in the same connection it is relevant that the measurements of Corliss, Delabarre and Elliott (1950) on the powder susceptibility of MnF_2 above 180°K, i.e. above $2{\cdot}5T_c$, lead to $\theta = 97°$K which gives $\theta/T_c = 1{\cdot}35$. While Li's values of θ/T_c and θ_{lim}/T_c may be somewhat altered in going from $S = 1/2$ to $S = 5/2$ these results make it seem unlikely that in either MnF_2 or FeF_2 is there any appreciable interaction between atoms on the same sub-lattice, i.e. along the tetragonal c axis. This conclusion does not of course hold for compounds such as MnO, MnS and FeO where the observed ratios are very much larger (5·0, 3·2 and 2·9 respectively).

The Bethe method, although it provides a satisfactory description of the Ising model at all temperatures, unfortunately breaks down at low temperatures when applied to the Heisenberg–Van Vleck model. In addition to a normal Curie point which marks the appearance of order as the system is cooled, there also occurs a lower critical temperature ('anti-Curie point') at which order disappears again. This failure of the Bethe–Weiss method has been discussed by P. W. Anderson (1950c). A similar situation results if we try to use the Kirkwood (1938) moment expansion method to evaluate the partition function (Heisenberg 1928, Tredgold 1953).

§5. The Transition Metals

The qualitative success of the Heisenberg (1928) theory of ferromagnetism in Fe, Co and Ni suggests that the other 3d transition metals may be regarded as antiferromagnetics, and for this reason we conclude with a brief survey of their present status. Unfortunately, however, the Heitler–London–Heisenberg approach is far less illuminating when applied to the metals than when applied to their compounds. The questions that have to be answered experimentally are largely questions of the fundamental applicability of this picture, e.g. are there any anomalies in the specific heats of, say, Cr and Mn and could these indicate the disappearance of antiferromagnetic order? When Néel (1936b) published the original molecular field treatment of antiferromagnetism, this appeared to be an adequate explanation of the specific heat peak found in α-manganese at about

375°c by Ashworth (1936). More recently Armstrong and Grayson-Smith (1950) have again studied the specific heat of manganese (99·99% purity) but were unable to find any anomalous behaviour in the range of temperature covered by their experiments (0°c to 800°c), except for that due to the known α–β transition at 717°c. Similarly they found no peak in the specific heat of chromium between 0°c and 800°c and it has since been confirmed that no anomalies present themselves below 0°c (Goldman 1953). In view of these facts and the suggestion (Armstrong and Grayson-Smith 1950, McGuire and Kriessman 1952) that the antiferromagnetic Curie points may lie at very high temperatures, one must attach the greatest importance to the neutron diffraction experiments which have been performed recently by Shull and Wilkinson (1953) on the transition elements W, Mo, Nb, V, Cr, Mn and Fe.

The (outer) electronic configurations of the free atoms of these elements are shown in table 2. If these configurations were also representative of the

Table 2. Electronic Configurations existing in Isolated Atoms of various Transition Elements

	3d	4s	4d	5s	5d	6s
V	3	2				
Cr	5	1				
Mn	5	2				
Fe	6	2				
Nb			4	1		
Mo			5	1		
W					4	2

metallic atoms, we would expect to find large atomic moments in the metals, since Hund's rule tells us that the spins of the d-electrons will be aligned parallel to one another as far as possible. This expectation still holds even if, as Zener (1951) has suggested, some of the atomic s-electrons become 'demoted' to the d-shell in the metal in order to minimize the Fermi energy and so increase the metallic binding. Such large atomic moments should contribute (*a*) a pronounced coherent scattering to the neutron diffraction pattern if there is any correlation in the alignment of neighbouring spins or (*b*) a measurable diffuse scattering if the moments are uncorrelated (Halpern and Johnson 1939). However the data obtained by Shull and Wilkinson for W, Nb, Mo and V indicate the complete absence of any measurable coherent or incoherent magnetic scattering, either at 20°K or at room temperature. In addition vanadium and niobium were studied at liquid helium temperature (where they are superconducting) but again with negative results. These facts, which are difficult to reconcile with the Néel–Zener approach, are in fact to be expected on the basis of the collective electron theory. In this theory the d electrons as well as the conduction electrons (s) are described by non-localized wave functions; in other words they are shared collectively by the whole metal lattice. In paramagnetic metals the ion cores will therefore be surrounded as often by electrons of antiparallel spin as by electrons of parallel spin, and the net moment of each atom is zero. It has been shown by Kriessman (1953) that this approach can also account for the different ways in which the observed susceptibilities vary with temperature.

On the other hand the neutron diffraction data for chromium and manganese provide evidence of a more complicated situation. In chromium there is a weak

super-lattice (100) reflection whose intensity depends on temperature as shown in fig. 23 (cf. fig. 14). This shows the presence of antiferromagnetic order below about 475°K. From the intensity at absolute zero (0·0038 barn, steradian per Cr atom) one can evaluate the magnetic moment located on each Cr atom; this is found to be about $0{\cdot}4\beta$, a value very much less than that expected on Zener's model. Shull and Wilkinson show in addition that there is no incoherent paramagnetic scattering to account for the difference. Super-lattice reflections are also found in α-manganese below a Curie temperature equal to about 100°K. However the complicated lattice of the α-form (see Seitz 1940, p. 8) does not permit an unambiguous determination of the atomic moment, although the indications are that it is about $0{\cdot}5\beta$.

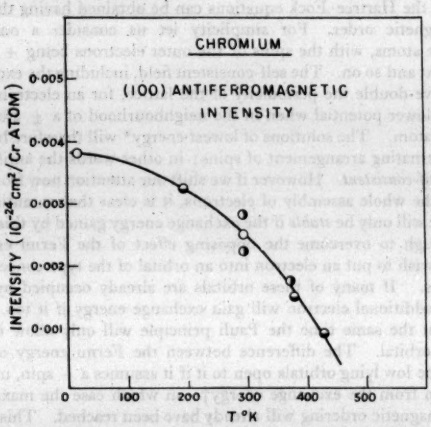


Fig. 23. Temperature dependence of the Cr (100) antiferromagnetic intensity in absolute units of differential scattering cross section (*after* Shull and Wilkinson 1953).

Although chromium and manganese have thus been shown to display antiferromagnetic ordering, their small moments are almost as difficult for the Zener theory to accommodate as the zero moments of V, Nb, Mo and W (see however Zener and Heikes 1953). Apart from this it would be unsatisfactory to be obliged to use the collective electron theory for vanadium and a Heitler–London theory for chromium—its neighbour in the periodic table. A promising solution is provided by a recent extension of the collective electron theory due to Slater* (1951a). Although this does not enable us to predict in advance which metals will be antiferromagnetic, it offers hope of being able to handle ferromagnetic

* Slater's paper was written mainly with a view to antiferromagnetic compounds, in an attempt to explain why they are not conductors. We do not feel that this application is justified, because although his arguments appear to show that substances such as MnO will be insulators at 0°K, it is impossible to escape the conclusion that they will be metallic conductors above their Curie points. A reformulation of the problem by the author (to be published) suggests that the antiferromagnetic electrons will actually conduct at all temperatures.

(e.g. Ni), antiferromagnetic (e.g. Cr) and simple paramagnetic (e.g. V) metals all together in one scheme.

Slater begins by regarding the one-electron wave functions as solutions of the Hartree–Fock equations which, as he himself had shown earlier (Slater 1951 b), can be interpreted as ordinary Schrödinger equations. On the basis of this interpretation we can say that each electron moves in a potential which is made up of the (average) electrostatic fields of the ion cores and of the other electrons, reduced by an 'exchange charge correction'. This exchange correction is such that the potential of a given electron is lower when it is in a region containing other electrons of the same spin, than when it is in a region containing electrons of the opposite spin. On this basis we can easily see how self-consistent solutions of the Hartree–Fock equations can be obtained having the character of antiferromagnetic order. For simplicity let us consider a one-dimensional array of like atoms, with the spins of the outer electrons being + on one atom, − on the next and so on. The self-consistent field, including the exchange charge, will now have double the periodicity of the lattice, for an electron with $\pm$ spin will be at a lower potential when in the neighbourhood of a $\pm$ atom than when near to a $\mp$ atom. The solutions of lowest energy* will therefore be those which keep the alternating arrangement of spins; in other words the antiferromagnetic solution is *self-consistent*. However if we shift our attention now from a particular electron to the whole assembly of electrons, it is clear that an antiferromagnetic arrangement will only be *stable* if the exchange energy gained by this arrangement is large enough to overcome the opposing effect of the Fermi energy. Thus suppose we wish to put an electron into an orbital of the type concentrated about the + atoms. If many of these orbitals are already occupied by electrons of + spin the additional electron will gain exchange energy if it too takes on a + spin. But at the same time the Pauli principle will only allow it to enter an unoccupied orbital. The difference between the Fermi energy of this orbital and that of the low lying orbitals open to it if it assumes a − spin, may be greater than the gain from the exchange energy†—in which case the maximum degree of antiferromagnetic ordering will already have been reached. This model therefore quite naturally allows small and non-integral values for the net atomic moment in antiferromagnetic metals. Quantitative statistical calculations have yet to be made for this model, but they will obviously be formally the same as those for the collective electron model of ferromagnetism (Stoner 1938, 1939, Wohlfarth 1951). Indeed we can make use of this identity to understand an apparent inconsistency in the experimental data. Although the Curie point of chromium is 200°c Armstrong and Grayson-Smith, as we have already mentioned, found no evidence of any anomalous specific heat at this temperature. Now Stoner (1939, especially table IV) using the collective electron model has calculated the heights ΔC of the specific heat peaks in ferromagnetics as a function of their relative magnetization at absolute zero, ζ_0. This quantity is the ratio of the

* The solutions for a $\pm$ spin are divided into two groups, and the corresponding one-electron Hartree–Fock energies are divided into two bands with a region of forbidden energy between them. The low energy band corresponds to solutions for $\pm$ spin which are concentrated around only the $\pm$ atoms: the high energy band for $\pm$ spin corresponds to solutions which are concentrated only about the $\mp$ atoms.

† We are implicitly decomposing the Hartree–Fock one-electron energy into two parts (i) the exchange energy and (ii) the Fermi energy, by which we mean kinetic energy and coulomb energy of interaction with the lattice and with other electrons.

Table 3

Note.—So far as possible we have selected the experimental papers listed here as being the most recent and most complete. This does not imply that they are of comparable reliability; in particular the specific heat data obtained by Trapeznikowa *et al.* (1936) on $NiCl_2$ are known to be unreliable (Busey and Giauque 1952), and this casts doubt on the accuracy of the results reported at the same time for $CoCl_2$, although there appear to be no later studies of this latter compound. We have listed only those compounds which one may judge (with differing degrees of certainty) to be antiferromagnetic on the basis of measurements already made. We have not included compounds (e.g. VCl_2) which would possibly turn out to be antiferromagnetic if measurements were made at lower temperatures than hitherto.

Compound	Magnetic susceptibility	Magnetic anisotropy	Paramagnetic resonance	Neutron diffraction	Specific heat	Thermal dilatation	Thermal changes in lattice parameters
$CoCl_2$	Shalyt 1938 Starr, Bitter & Kaufmann 1940				Trapeznikowa, Schubnikow & Miljutin 1936 Miljutin & Shalyt 1939		
CoF_2	Bizette 1946	Stout & Matarrese 1953		Erickson 1953	Stout & Catalano 1953		
CoI_2	Shalyt 1939 b						
CoO	Henry la Blanchetais 1951		Maxwell & McGuire 1953	Shull, Strauser & Wollan 1951		M. Foëx 1948	Greenwald & Smart 1950
$CrCl_2$	Starr, Bitter & Kaufmann 1940				C. T. Anderson 1937		
Cr_2O_3	G. Foëx & Graff 1939 Volger 1952		Maxwell & McGuire 1953	Brockhouse 1953	C. T. Anderson 1937 Volger 1952		Greenwald 1951
CrSb	Haraldsen, Rosenqvist & Grønvold 1949			Snow 1953			Snow 1953 Willis 1953
$CuCl_2$	Starr, Bitter & Kaufmann 1940						
$CuCl_2.2H_2O$	Van den Handel, Gijsman & Poulis 1952		Ubbink, Poulis, Gerritsen & Gorter 1952		Friedberg 1952		
CuO	Klemm & Schüth 1931		Maxwell & McGuire 1953				
$FeCl_2$	Starr, Bitter & Kaufmann 1940 Shalyt 1939 b				Trapeznikowa & Schubnikow 1935 Miljutin & Shalyt 1939		
$FeCO_3$	Bizette & Tsai 1949	G. Foëx 1921			C. T. Anderson 1934		
FeF_2	Bizette & Tsai	Stout & Matarrese		Erickson 1953	Stout & Catalano		
FeO	Bizette & Tsai 1943			Shull, Strauser & Wollan 1951	Millar 1929	M. Foëx 1948	Tombs & Rooksby 1950 Willis & Rooksby 1953

								Willis & Rooksby 1953
α-Fe_2O_3	Bizette, Squire & Tsai 1938			Shull, Strauser & Wollan 1951				Willis & Rooksby 1952
Fe_2SiO_4	Chevallier 1951				Kelley 1941			
$MnCl_2$	Starr, Bitter & Kaufmann 1940 Erickson*				Trapeznikowa & Miljutin 1937			
$MnCl_2.4H_2O$	Henry 1953				Friedberg & Wasscher 1954			
MnF_2	de Haas, Schultz & Koolhaas 1940	Griffel & Stout 1950		Erickson 1953	Stout & Adams 1942			
MnO	Bizette, Squire & Tsai 1938		Maxwell & McGuire 1953	Shull, Strauser & Wollan 1951	Millar 1928	M. Foëx 1948		Tombs & Rooksby 1950
MnO_2	Bizette & Tsai 1949	Bizette & Tsai 1949		Erickson 1952	Kelley & Moore 1943			
MnS	Bizette 1946 Serres 1947		Maxwell & McGuire 1953	Shull, Strauser & Wollan 1951	C. T. Anderson 1931			Rooksby & Tombs 1951
MnSe	Lindsay 1951		Maxwell & McGuire 1953	Shull, Strauser & Wollan 1951	Kelley 1939			
MnTe	Squire 1939 Serres 1947				Kelley 1939			Greenwald 1953
NiF_2	Bizette 1946			Erickson 1953	Stout & Catalano 1953			
$NiCl_2$	Starr, Bitter & Kaufmann 1940				Busey & Giauque 1952			
NiO	M. Foëx & Henry la Blanchetais 1949			Shull, Strauser & Wollan 1951		M. Foëx 1948		Rooksby 1948
VCl_2	Starr, Bitter & Kauffman 1940							
V_2O_3	Hoschek & Klemm 1939 G. Foëx & Wucher 1949				C. T. Anderson 1936 Jaffray & Lyand 1951	M. Foëx 1946, 1949		
VO_2	Perakis & Wucher 1952				C. T. Anderson 1936			

* Private communication (1953) indicating that $T_c \simeq 2°K$.

net moment per atom to the maximum possible moment per atom. In chromium, if we accept Shull and Wilkinson's value for the net moment, we have ζ_0 equal to 0·4/5, i.e. $\sim$0·1. Stoner's calculations show that when ζ_0 is as small as this, ΔC is too small to be observable. The measurements of McGuire and Kriessman (1952) on chromium seem to show that neither is there any appreciable anomaly in the magnetic susceptibility, though we cannot infer this from the calculations for the ferromagnetic case.

Note added in proof (22.12.53). Calculations along the lines indicated here have since been made by the author and are in encouraging agreement with experiment. A preliminary account (Lidiard 1953) contains formulae for the susceptibility $\chi_{\parallel}$; the perpendicular susceptibility $\chi_{\perp}$ is independent of temperature below the Curie point and equal to $\chi_{\parallel}(T_c)$ as is the case with the Van Vleck model. A fuller discussion is in course of publication.

Acknowledgments

This Report was written while I was at the University of Illinois; it is a pleasure to express my thanks to Professor F. Seitz and his colleagues at Urbana for their generous hospitality and encouragement. I would also like to acknowledge the benefit of many helpful discussions with Dr. Y. Yafet. Lastly thanks are due to my wife for her assistance with the manuscript and to those authors who have sent reprints and reports of their work.

References

Abragam, A., and Pryce, M. H. L., 1951, *Proc. Roy. Soc.* A, **205,** 135.

Anderson, C. T., 1931, *J. Amer. Chem. Soc.*, **53,** 476; 1934, *Ibid.*, **56,** 849; 1936, *Ibid.*, **58,** 564; 1937, *Ibid.*, **59,** 488.

Anderson, P. W., 1950 a, *Phys. Rev.*, **79,** 350; 1950 b, *Ibid.*, **79,** 705; 1950 c, *Ibid.*, **80,** 922; 1951, *Ibid.*, **83,** 1260; 1952, *Ibid.*, **86,** 694.

Armstrong, L. D., and Grayson-Smith, H., 1950, *Canad. J. Res.* A, **28,** 51.

Ashworth, J. R., 1936, *Proc. Phys. Soc.*, **48,** 456.

Bacon, G. E., and Lonsdale, K., 1953, *Rep. Progr. Phys.*, **16,** 1 (London : Physical Society).

Bacon, G. E., Street, R., and Tredgold, R. H., 1953, *Proc. Roy. Soc.* A, **217,** 252.

Bacher, R. F., and Goudsmit, S., 1932, *Atomic Energy States* (New York : McGraw-Hill).

Bethe, H. A., 1931, *Z. Phys.*, **71,** 205.

Bitter, F., 1938, *Phys. Rev.*, **54,** 79.

Bizette, H., 1946, *Ann. Phys., Paris*, (12), **1,** 295; 1951, *J. Phys. Radium*, **12,** 161.

Bizette, H., Squire, C. F., and Tsai, B., 1938, *C. R. Acad. Sci., Paris*, **207,** 449.

Bizette, H., and Tsai, B., 1939, *C. R. Acad. Sci., Paris*, **209,** 205; 1941 a, *Ibid.*, **212,** 75; 1941 b, *Ibid.*, **212,** 119; 1943, *Ibid.*, **217,** 390; 1949, *Polarisation de la matière* (Paris : Centre National de la Recherche Scientifique), p. 164.

Bleaney, B., and Bowers, K. D., 1952, *Phil. Mag.*, **43,** 372.

Bleaney, B., and Stevens, K. W. H., 1953, *Rep. Progr. Phys.*, **16,** 108 (London : Physical Society).

Bloch, F., 1930, *Z. Phys.*, **61,** 206; 1932, *Ibid.*, **74,** 295.

Brockhouse, B. N., 1953, *J. Chem. Phys.*, **21,** 961.

Busey, R. H., and Giauque, W. F., 1952, *J. Amer. Chem. Soc.*, **74,** 4443.

Chevallier, R., 1951, *J. Phys. Radium*, **12,** 172.

Corliss, L., Delabarre, Y., and Elliott, N., 1950, *J. Chem. Phys.*, **18,** 1256.

Corson, E. M., 1951, *Perturbation Methods in the Quantum Mechanics of n-Electron Systems* (London : Blackie).

Coulson, C. A., 1952, *Valence* (Oxford : Clarendon Press).

Dirac, P. A. M., 1947, *Quantum Mechanics*, 3rd Edn. (Oxford : Clarendon Press).

ELLEFSON, B. S., and TAYLOR, N. W., 1934, *J. Chem. Phys.*, **2**, 58.
ELLIOTT, N., and WOLFSBERG, M., 1953, *Phys. Rev.*, **91**, 435.
ERICKSON, R. A., 1952, *Phys. Rev.*, **85**, 745; 1953, *Ibid.*, **90**, 779.
ERICKSON, R. A., and SHULL, C. G., 1951, *Phys. Rev.*, **83**, 208.
FAIRWEATHER, A., ROBERTS, F. F., and WELCH, A. J. E., 1952, *Rep. Progr. Phys.*, **15**, 142 (London : Physical Society).
FINE, M. E., 1953, *Rev. Mod. Phys.*, **25**, 158.
FOËX, G., 1921, *Ann. Phys., Paris*, (9), **16**, 255.
FOËX, G., and GRAFF, M., 1939, *C. R. Acad. Sci., Paris*, **209**, 160.
FOËX, G., KARANTASSIS, T., and PERAKIS, N., 1953, *C. R. Acad. Sci., Paris*, **237**, 982.
FOËX, G., and WUCHER, J., 1949, *C. R. Acad. Sci., Paris.*, **229**, 882.
FOËX, M., 1946, *C. R. Acad. Sci., Paris*, **223**, 1126; 1948, *Ibid.*, **227**, 193; 1949, *Ibid.*, **229**, 880.
FOËX, M., and HENRY LA BLANCHETAIS, C., 1949, *C. R. Acad. Sci., Paris*, **228**, 1579.
FOWLER, R. H., and GUGGENHEIM, E. A., 1949, *Statistical Thermodynamics* (Cambridge : University Press).
FRIEDBERG, S. A., 1952, *Physica*, **18**, 714.
FRIEDBERG, S. A., and WASSCHER, J. D., 1954, *Physica*, in the press.
GOLDMAN, J. E., 1953, *Rev. Mod. Phys.*, **25**, 113.
GORTER, C. J., 1953 a, *Rev. Mod. Phys.*, **25**, 277; 1953 b, *Ibid.*, **25**, 332.
GORTER, C. J., and HAANTJES, J., 1952, *Physica*, **18**, 285.
GREENWALD, S., 1951, *Nature, Lond.*, **168**, 379; 1953, *Acta Cryst., Camb.*, **6**, 396.
GREENWALD, S., and SMART, J. S., 1950, *Nature, Lond.*, **166**, 523.
GRIFFEL, M., and STOUT, J. W., 1950, *J. Chem. Phys.*, **18**, 1455.
GUGGENHEIM, E. A., 1952, *Mixtures* (Oxford : University Press).
DE HAAS, W. J., SCHULTZ, B. H., and KOOLHAAS, J., 1940, *Physica*, **7**, 57.
HALPERN, O., and JOHNSON, M. H., 1939, *Phys. Rev.*, **55**, 898.
HARALDSEN, H., ROSENQVIST, T., and GRØNVOLD, F., 1949, *Arch. Math. Naturvidensk.*, **50**, 95.
HARRISON, R. J., 1950, *M.I.T. Research Laboratory of Electronics*, Technical Report No. 172.
HARTREE, D. R., 1947, *Rep. Progr. Phys.*, **11**, 113 (London : Physical Society).
HEISENBERG, W., 1928, *Z. Phys.*, **49**, 619.
HELLER, G., and KRAMERS, H. A., 1934, *Proc. Roy. Acad. Sci., Amsterdam*, **37**, 378.
HENRY, W. E., 1953, *Phys. Rev.*, **90**, 492.
HENRY LA BLANCHETAIS, C., 1951, *J. Phys. Radium*, **12**, 765.
HOLSTEIN, T., and PRIMAKOFF, H., 1940, *Phys. Rev.*, **58**, 1098.
HOSCHEK, E., and KLEMM, W., 1939, *Z. anorg. allg. Chemie*, **242**, 63.
HOWARD, L. N., and SMART, J. S., 1953, *Phys. Rev.*, **91**, 17.
HULTHÉN, L., 1936, *Proc. Roy. Acad. Sci., Amsterdam*, **39**, 190.
ISING, E., 1925, *Z. Phys.*, **31**, 253.
JAFFRAY, J., and LYAND, R., 1951, *C. R. Acad. Sci., Paris*, **233**, 133.
JOST, W., 1952, *Diffusion in Solids, Liquids, Gases* (New York : Academic Press).
KAPLAN, H., 1952, *Phys. Rev.*, **85**, 1038.
KEFFER, F., 1952, *Phys. Rev.*, **87**, 608.
KEFFER, F., KAPLAN, H., and YAFET, Y., 1953, *Amer. J. Phys.*, **21**, 250.
KELLEY, K. K., 1939, *J. Amer. Chem. Soc.*, **61**, 203; 1941, *Ibid.*, **63**, 2750.
KELLEY, K. K., and MOORE, G. E., 1943, *J. Amer. Chem. Soc.*, **65**, 782.
KIRKWOOD, J. G., 1938, *J. Chem. Phys.*, **6**, 70.
KITTEL, C., 1951, *Phys. Rev.*, **82**, 565.
KLEMM, W., and SCHÜTH, W., 1931, *Z. anorg. allg. Chemie*, **203**, 104.
KRAMERS, H. A., 1934, *Physica*, **1**, 182; 1952, *Ibid.*, **18**, 101.
KRIESSMAN, C. J., 1953, *Rev. Mod. Phys.*, **25**, 122.
KRISHNAN, K. S., and BANERJEE, S., 1936, *Phil. Trans. Roy. Soc.* A, **235**, 343.
KRISHNAN, K. S., CHAKRAVORTY, N. C., and BANERJEE, S., 1933, *Phil. Trans. Roy. Soc.* A, **232**, 99.
KUBO, R., 1952, *Phys. Rev.*, **87**, 568.
LI, Y.-Y., 1951, *Phys. Rev.*, **84**, 721.
LIDIARD, A. B., 1953, *Proc. Phys. Soc.* A, **66**, 1188.
LINDSAY, R., 1951, *Phys. Rev.*, **84**, 569.

McGuire, T. R., and Kriessman, C. J., 1952, *Phys. Rev.*, **85**, 452.
Maxwell, L. R., and McGuire, T. R., 1953, *Rev. Mod. Phys.*, **25**, 279.
Michel, A., Chaudron, G., and Bénard, J., 1951, *J. Phys. Radium*, **12**, 189.
Miljutin, G. A., and Shalyt, S. S., 1939, *C. R. Acad. Sci., U.S.S.R.*, **24**, 680.
Millar, R. W., 1928, *J. Amer. Chem. Soc.*, **50**, 1875; 1929, *Ibid.*, **51**, 215.
Mott, N. F., and Gurney, R. W., 1948, *Electronic Processes in Ionic Crystals* (Oxford : University Press).
Nagamiya, T., 1951, *Progr. Theor. Phys.*, **6**, 342.
Néel, L., 1932, *Ann. Phys., Paris*, (10), **18**, 5; 1936 a, *Ibid.*, (11), **5**, 232; 1936 b, *C. R. Acad. Sci., Paris*, **203**, 304; 1948, *Ann. Phys., Paris*, (12), **3**, 137.
Newell, G. F., and Montroll, E. W., 1953, *Rev. Mod. Phys.*, **25**, 353.
Nix, F. C., and Shockley, W., 1938, *Rev. Mod. Phys.*, **10**, 1.
Opechowski, W., 1937, *Physica*, **4**, 181; 1938, *Ibid.*, **6**, 1112.
Penney, W. G., 1937, *Proc. Roy. Soc.* A, **158**, 306.
Perakis, N., and Wucher, J., 1952, *C. R. Acad. Sci., Paris*, **235**, 354.
Polder, D., 1951, *J. Phys. Radium*, **12**, 274.
Poulis, N. J., Van den Handel, J., Ubbink, J., Poulis, J. A., and Gorter, C. J., 1951, *Phys. Rev.*, **82**, 552.
Pryce, M. H. L., 1950, *Proc. Phys. Soc.* A, **63**, 25.
Rooksby, H. P., 1948, *Acta Cryst., Camb.*, **1**, 226.
Rooksby, H. P., and Tombs, N. C., 1951, *Nature, Lond.*, **167**, 364.
Seitz, F., 1940, *Modern Theory of Solids* (New York : McGraw–Hill).
Serres, A., 1947, *J. Phys. Radium*, **8**, 146.
Shalyt, S. S., 1938, *J. Exp. Theor. Phys.*, **8**, 518; 1939 a, *Nature, Lond.*, **143**, 799; 1939 b, *J. Exp. Theor. Phys.*, **9**, 1073.
Shull, C. G., and Smart, J. S., 1949, *Phys. Rev.*, **76**, 1256.
Shull, C. G., Strauser, W. A., and Wollan, E. O., 1951, *Phys. Rev.*, **83**, 333.
Shull, C. G., and Wilkinson, M. K., 1953, *Rev. Mod. Phys.*, **25**, 100.
Slater, J. C., 1929, *Phys. Rev.*, **34**, 1293; 1951 a, *Ibid.*, **82**, 538; 1951 b, *Ibid.*, **81**, 385; 1953, *Rev. Mod. Phys.*, **25**, 199.
Smart, J. S., 1952, *Phys. Rev.*, **86**, 968; 1953 a, *Rev. Mod. Phys.*, **25**, 327; 1953 b, *Phys. Rev.*, **90**, 55.
Smart, J. S., and Greenwald, S., 1951, *Phys. Rev.*, **82**, 113.
Snow, A. I., 1952, *Phys. Rev.*, **85**, 365; 1953, *Rev. Mod. Phys.*, **25**, 127.
Squire, C. F., 1939, *Phys. Rev.*, **56**, 922.
Starr, C., Bitter, F., and Kaufmann, A. R., 1940, *Phys. Rev.*, **58**, 977.
Stoner, E. C., 1938, *Proc. Roy. Soc.* A, **165**, 372; 1939, *Ibid.*, **169**, 339; 1947, *Rep. Progr. Phys.*, **11**, 43 (London : Physical Society).
Stout, J. W., and Adams, H. E., 1942, *J. Amer. Chem. Soc.*, **64**, 1535.
Stout, J. W., and Catalano, E., 1953, *Phys. Rev.*, **92**, 1575.
Stout, J. W., and Matarrese, L. M., 1953, *Rev. Mod. Phys.*, **25**, 338.
Street, R., 1951, *Sci. Progr.*, **39**, 258.
Street, R., and Lewis, B., 1951, *Nature, Lond.*, **168**, 1036.
Tessman, J. R., 1952, *Phys. Rev.*, **88**, 1132.
Tombs, N. C., and Rooksby, H. P., 1950, *Nature, Lond.*, **165**, 442.
Trapeznikowa, O. N., and Miljutin, G. A., 1937, *Phys. Z. Sowjet*, **11**, 55.
Trapeznikowa, O. N., and Schubnikow, L. W., 1935, *Phys. Z. Sowjet*, **7**, 66.
Trapeznikowa, O. N., Schubnikow, L. W., and Miljutin, G. A., 1936, *Phys. Z. Sowjet*, **9**, 237.
Tredgold, R. H., 1953, *Proc. Phys. Soc.* A, **66**, 421.
Ubbink, J., 1953, *Physica*, **19**, 9.
Ubbink, J., Poulis, J. A., Gerritsen, H. J., and Gorter, C. J., 1952, *Physica*, **18**, 361.
Van den Handel, J., Gijsman, H. M., and Poulis, N. J., 1952, *Physica*, **18**, 862.
Van Vleck, J. H., 1932, *The Theory of Electric and Magnetic Susceptibilities* (Oxford : University Press); 1940, *Le Magnétisme*, **3**, 103 (Paris : Centre National de la Recherche Scientifique); 1941, *J. Chem. Phys.*, **9**, 85; 1945, *Rev. Mod. Phys.*, **17**, 27 ; 1951, *J. Phys. Radium*, **12**, 262.
Van Vleck, J. H., and Penney, W. G., 1934, *Phil. Mag.*, **17**, 961.
Volger, J., 1952, *Nature, Lond.*, **170**, 1027.

WANNIER, G. H., 1950, *Phys. Rev.*, **79**, 357.
WEISS, P. R., 1948, *Phys. Rev.*, **74**, 1493.
WILLIS, B. T. M., 1953, *Acta Cryst., Camb.*, **6**, 425.
WILLIS, B. T. M., and ROOKSBY, H. P., 1952, *Proc. Phys. Soc.* B, **65**, 950; 1953, *Acta Cryst., Camb.*, **6**, 827.
WOHLFARTH, E. P., 1949, *Nature, Lond.*, **163**, 57; 1951, *Phil. Mag.*, **42**, 374.
WYCKOFF, R. W. G., 1948– , *Crystal Structures* (New York : Interscience).
YOSIDA, K., 1951 a, *Progr. Theor. Phys.*, **6**, 691; 1951 b, *Ibid.*, **6**, 356; 1952. *Ibid.*, **7**, 25.
ZEHLER, V., 1950, *Z. Naturforsch.*, **5 a**, 344.
ZENER, C., 1951, *Phys. Rev.*, **81**, 440.
ZENER, C., and HEIKES, R. R., 1953, *Rev. Mod. Phys.*, **25**, 191.
ZIMAN, J. M., 1952, *Proc. Phys. Soc.* A, **65**, 540, 548; 1953, *Ibid.*, **66**, 89.

HORIZONTAL MOVEMENTS IN THE IONOSPHERE

By B. H. BRIGGS and M. SPENCER
Cavendish Laboratory, Cambridge

CONTENTS

	PAGE
§1. Introduction	246
§2. Reflected radio waves observed at closely spaced receiving points	246
2.1. Description of method	246
2.2. Results	249
2.2 (*a*). Magnitude	249
2.2 (*b*). Diurnal and seasonal changes	250
2.2 (*c*). Relation to magnetic activity	262
2.2 (*d*). Long period changes	264
2.2 (*e*). Short period fluctuations	265
§3. Observations of movements of large scale irregularities in the F region	265
3.1. Description of methods	265
3.2. Results	267
§4. Observations of waves from radio stars	269
4.1. Description of method	269
4.2. Results	270
§5. Observations of waves reflected from meteor trails	271
5.1. Description of method	271
5.2. Results	272
§6. Observations of waves reflected from the sporadic-E layer	275
§7. Summary and conclusions	276
7.1. E region (closely spaced receivers; meteor observations)	276
7.2. F region (closely spaced receivers; movements of large irregularities; observations by radio astronomy)	277
7.3. Non-radio methods	278
Acknowledgments	279
References	280

Abstract. The object of the paper is to survey the existing knowledge of horizontal movements in the ionosphere as obtained by radio methods. After a brief introduction the various methods and associated results are described under the following main headings: Reflected radio waves observed at closely spaced points; Observations of movements of large scale irregularities in the F region; Observations of waves from radio stars; Observations of waves reflected from meteor trails; Observations of waves reflected from the sporadic-E layer. The most detailed discussion is accorded to the first of these methods, with which the authors have been personally associated. The results obtained at Cambridge, England, by this method are compared with the results obtained by other workers using this and the other methods in different parts of the world. The following conclusions are reached: At all levels in the ionosphere systematic as well as irregular horizontal movements are present. Typical velocities are of the order of 80 m/sec; there is a tendency for the velocity to increase with height. The systematic movements have the form of regular diurnal and seasonal variations and show considerable consistency between observations made in the Northern and Southern hemispheres. At heights around 100 km the semi-diurnal components of the velocity, as determined by Fourier analysis, are found to represent a uniformly rotating velocity vector; the direction of the rotation is clockwise in the Northern hemisphere and anti-clockwise in the Southern hemisphere. This is the type of movement predicted by the theory of atmospheric

oscillations, though the phase of the rotation is not that expected theoretically. At times of magnetic disturbance the velocities of the horizontal movements are increased, and this phenomenon becomes more marked with increasing height in the atmosphere. At 300 km velocities as high as 1000 m/sec have been recorded during magnetic storms. The report ends with a brief summary of observations by non-radio methods.

§1. Introduction

In recent years several different radio techniques have been developed for measuring the velocities of horizontal movements taking place in the ionosphere. It is the purpose of this paper to describe these methods in outline and to discuss the results obtained.

All the methods depend upon the existence of irregularities in the ionization which can be detected and followed. Most methods make use of signals emitted by transmitters on the ground and reflected from the ionosphere. One method makes use of the radio waves emitted by radio stars.

We have ourselves been associated with a group of workers who have used a method in which the fading of a wave reflected from the ionosphere is observed at closely spaced receiving points. This method has been used at Cambridge, England, over a period of years. We shall discuss the results of this method first, and most fully. We shall then compare and contrast the results obtained by other workers, using this and other methods, so far as we know of them. We are much indebted to the workers concerned for allowing us to use their results which, in some cases, have not yet been published.

The directions of movement given will be the direction *towards which* the motion is directed. It was agreed at the Xth General Assembly of U.R.S.I., held in Sydney in 1951, that this convention should be used by workers in this subject; it is the opposite to the convention used by meteorologists for wind directions. Velocities will be given in metres/sec. Times will be given as local time at the place of observation.

We shall not attempt to discuss here the important but difficult question as to whether the movements detected by the various radio techniques are true movements of the air. It may be that they are movements of the ionization only, without associated movement of the air. In some cases the movements may correspond to travelling disturbances or wave motions in the ionosphere, and not to bodily movements of electrons. It seems that evidence on this question may be obtained in due course from a comparison of the results of the various methods. In order to leave the question open, we shall not speak of 'winds' but only of 'movements'.

§2. Reflected Radio Waves observed at Closely Spaced Receiving Points

2.1. *Description of Method*

In the method of closely spaced receivers the amplitude of a wave reflected from the ionosphere is recorded at several points separated by distances of the order of one wavelength or less. The amplitude is found to vary in a random way and, when the 'fading' records obtained at the different points are compared, it is often found that they are similar except for the presence of time shifts. This effect was first reported by Pawsey (1935), and the method was developed specifically for the observation of movements in the ionosphere by Mitra (1949), Krautkrämer (1950), Phillips (1952) and others. It has been shown that the

fluctuations observed at a fixed point arise mainly from a drift over the ground of an amplitude pattern. This pattern is produced by a process of diffraction at an irregular ionospheric layer (Booker, Ratcliffe and Shinn 1950). The velocity of movement of this pattern over the ground will be twice the horizontal velocity of the layer (Phillips 1952). The magnitude and direction of the ionospheric movement can be deduced from a measurement of the time shifts for a pair of receivers in a North–South line, and for a pair in an East–West line. The complications which arise if the diffraction pattern also changes as it moves, due to changes in the detailed structure of the ionospheric irregularities, have been discussed by Briggs, Phillips and Shinn (1950).

It is necessary to ensure that the amplitude recorded is that of a single downcoming wave. On the shorter wavelengths this is conveniently done by the usual echo technique, in which the echoes corresponding to the ground wave, the first reflection, and higher order reflections are separated in time. The echo corresponding to a single reflection is selected by a 'gate', and the amplitude of this echo alone is recorded. When the two magneto-ionic components are present with comparable amplitudes, part of the observed fading may be caused by their mutual interference and may lead to erroneous conclusions (Mitra 1949). To avoid this complication one of the magneto-ionic components can be removed by means of a circularly polarized aerial system (Phillips 1951). On longer wavelengths, where continuous wave transmissions must be used, the corresponding techniques become more difficult, and the most that can normally be done is to eliminate the ground wave by means of a directional loop.

The optimum distance between the receiving points in the method of closely spaced receivers is the greatest distance for which there is still considerable similarity between the fading records obtained. It is found that this distance is about one wavelength for waves of length around 50 to 500 metres (Briggs and Phillips 1950). For these wavelengths it is therefore convenient to place the receivers at a central point and to use spaced aerials connected to the receivers by radio-frequency feeders.

The fading curves for the different receivers may be recorded photographically or with pen recorders, and afterwards examined for time shifts. A more convenient method, in which the time shifts are recorded directly on a pen recorder, has been developed by Phillips.* It is arranged that two receivers, each fed from a separate aerial, produce d.c. voltages proportional to the amplitude of the selected echo. Circuits are arranged so that, for each voltage separately, a relay is closed just after a maximum value has been reached and is opened just after a minimum. It is further arranged that if minima occur at different times in the two receivers, a condenser is charged through a high resistance during the interval between the closing of the two relays, so that it acquires a charge proportional to the time interval between the minima. When the second relay closes, the condenser is discharged in such a way that an impulsive deflection, of magnitude proportional to the time delay, is produced on a pen recorder. The undeflected position of the pen is in the centre of the paper, and it is arranged that the sense of the deflection corresponds to the sense of the time shift.

In order to economize in apparatus, the time shifts for the two pairs of aerials, spaced along North–South and East–West lines, may be recorded alternately. Records obtained by this method are shown in fig. 1. In fig. 1(*a*)

* In course of publication.

the average time shifts for the two pairs of aerials are approximately constant, and from a knowledge of the aerial spacings it is possible to compute the magnitude and direction of the motion. In fig. 1 (*b*), however, the time shifts are variable in magnitude and direction, and it is not possible to deduce a steady motion over the period considered. The significance of this will be discussed in § 2.2 (*e*).

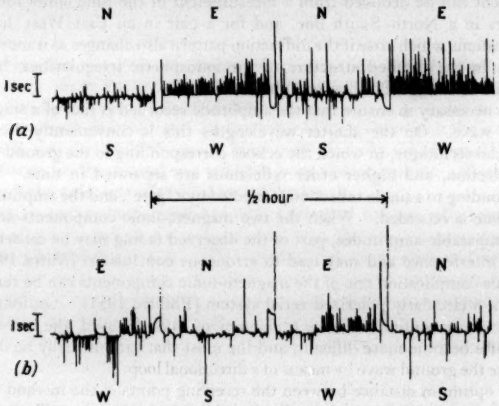


Fig. 1. Examples of records in which the time shifts at spaced receivers are recorded directly. The apparatus is switched every ½ hour from the EW pair to the NS pair. The records are labelled so that, for example, if the deflection is to the side marked E then a minimum in the signal on the eastern aerial follows one on the western. The time delay is given by the size of the deflections. (*a*) shows a well defined velocity towards the East; (*b*) shows irregular fluctuations of velocity.

The frequencies used by the various workers, and the period of the observations, are summarized in table 1.

Table 1. Observations by the Method of Closely Spaced Receivers

Krautkrämer, J. (1950)	3·5 Mc/s	Winter 1941–2.	Cologne.
Mitra, S. N. (1949)	2 to 6 Mc/s	Scattered observations during 1948.	Cambridge.
Phillips, G. J. (1952)	2·4 Mc/s	Feb. 1949–Sept. 1951.	Cambridge.
Spencer, M.	2 Mc/s & 4 Mc/s	Jan. 1952 onwards.	Cambridge.
Salzberg, C. D., and Greenstone, R. (1951)	2·3 Mc/s	March 1950 onwards.	Washington D.C. and Sterling, Virginia.
Chapman, J. H. (1953)	2 to 6 Mc/s	June 1950–Oct. 1950.	Ottawa.
		Nov. 1950–June 1951 and later.	Montreal.
Millman, G. H. (1952)	150 kc/s	July 1951–March 1952.	State College, Pennsylvania.
Bowhill, S. A. (1953)	70 kc/s	1951–2.	Cambridge.
Burtt, G. J.	2 Mc/s	Sept. 1952 onwards.	Lower Hutt, New Zealand.
Harang, L.	2 Mc/s	1953 onwards.	Kjeller, Norway.
Osborne, B. W.	6 Mc/s	1953 onwards.	Singapore, Malaya.

The method will give results for both the E region (height about 100 km) and the F region (200 to 300 km) if the frequency is chosen suitably. Several observers have used a single frequency around 2 Mc/s for which the results normally refer to the E region during the day and the F region at night. Occasional observations in the E region are possible during the night when sporadic-E ionization is present. With the lower frequencies of 70 kc/s and 150 kc/s, reflections from the E region are obtained by both day and night.

Results obtained with echoes reflected from the F region are open to doubt because the fading observed might be impressed on the wave by passage through the E region. The motion deduced would then refer to the E region, not the F region. We shall see, however, that the regular variations found for the F region are markedly different from those for the E region. This provides indirect evidence that the motions deduced from echoes returned from the F region probably refer to that region itself. It should be emphasized, however, that the uncertainty about the precise height of the measured movement is an inherent disadvantage of the present method. In principle, the irregularities causing the fading might be anywhere below the level of reflection. For this reason we shall simply divide the results into those appropriate to the E region and the F region, without attempting to specify the actual height any more precisely.

2.2. *Results*

The results obtained by the various workers who have used the method of closely spaced receivers are quite extensive, and it will be convenient to discuss them under the following sub-headings: (*a*) magnitude, (*b*) diurnal and seasonal changes, (*c*) relation to magnetic activity, (*d*) long period changes, (*e*) short period fluctuations.

2.2(*a*) *Magnitude.*

Krautkrämer (1950) gave histograms showing the frequency of occurrence of different values of velocity for the E region and the F region.* The values for the E region were in the range 25 m/sec to 200 m/sec, with a most frequent value of about 90 m/sec. For the F region velocities greater than 200 m/sec were sometimes obtained, but for velocities less than 100 m/sec the shape of the histogram was very similar to that for the E region, and the velocity most frequently observed was again about 90 m/sec.

Phillips (1952) gave histograms† in which results for the E and F regions were not separated; he did, however, distinguish between the values obtained in summer and winter. For summer the values were in the range 10 m/sec to 200 m/sec, with a most frequent value of 70 m/sec. For winter, the values were in the range 20 m/sec to 250 m/sec, with a most frequent value of about 100 m/sec. These results, therefore, suggest that on the average the velocity of movement is slightly greater in winter than in summer.

Salzberg and Greenstone (1951) gave separate histograms to show the frequency of occurrence of different values of velocity for ten different months in 1950. Results for E and F regions were not separated, and values greater than

* Krautkrämer's figures referred to the velocity measured at the ground; they have been halved here so as to represent the velocity in the reflecting layer.

† Unusually large velocities obtained for the F region, probably associated with magnetic disturbance, were not included in these histograms.

200 m/sec, which were only occasionally observed, were omitted. They stated that 70 m/sec was a representative speed for the greater part of the year, but that in November and December the speeds increased and were of the order of 100 m/sec.

Chapman (1953) gave separate histograms for observations using reflections from the E and F regions. For the E region the most frequent value was 80 m sec and for the F region 100 m/sec.

Millman (1952) used a lower frequency (150 kc/s) and obtained velocities ranging from 15 to 300 m/sec, with the most probable values between 40 and 90 m/sec. These results probably referred to heights of the order of 74 to 77 km in the daytime and 83 to 100 km at night. There appeared to be no great difference in the speeds measured by day and by night.

We may conclude from all these results that the velocity of movement most frequently observed is about 80 m/sec. There is some tendency for the values to be greater in winter than in summer, and greater for the F region than for the E region. Values greater than 200 m/sec are sometimes observed, especially in the F region, and appear to be connected with magnetic disturbance. This point will be discussed in § 2.2 (*c*).

2.2 (*b*) *Diurnal and seasonal changes.*

The velocity of movement is found to exhibit some regular diurnal and seasonal changes. Our discussion of these will be based on the series of observations made at Cambridge from February 1949 onwards by Phillips and Spencer, using frequencies in the 2 to 4 Mc/s range (table 1). Where appropriate, the results will be compared with those obtained by other workers using the same method.

Until September 1951 a single frequency of 2·4 Mc/s was used for most of the time, with occasional observations on higher frequencies. From January 1952 onwards two frequencies of 2 Mc/s and 4 Mc/s have been used simultaneously. During the day, waves with frequency of 2 Mc/s are reflected from the E region and 4 Mc/s from the F region. During the night, waves of both frequencies are reflected from the F region except when sporadic-E ionization is present. Results are available for the periods listed in table 2.

Table 2. Cambridge Observations

E region, day	1949–1953	2 Mc/s or 2·4 Mc/s
E region, night	1949–1953	Occasional sporadic-E reflections
		2 Mc/s and 4 Mc/s
F region, day	1952–1953	4 Mc/s
F region, night	1949–1953	2 Mc/s
	1952–1953	4 Mc/s

Observations were made continuously for three days every month. Figure 2 shows a plot of the results obtained for a typical month (October 1950). Each point represents an individual determination of the magnitude of the NS component or the EW component. The dotted curve was drawn by eye to represent the general trend of the points. From this curve average hourly values of velocity may be obtained, and these smoothed values were used in later analysis. This process removes the more rapid changes and irregular variations from one day to the next; we shall discuss these in § 2.2 (*e*).

The results shown in fig. 2 were obtained on a frequency of 2·4 Mc/s and refer to the E region by day and the F region by night. With the two frequencies in use from January 1952 separate mean curves could be drawn each month for the E and F regions. The mean curves obtained in this way for a particular month were

found to be similar for each of the years 1949 to 1953. The separate curves for each month were therefore averaged. As the twelve sets of curves cannot be reproduced here, we have condensed the data by giving curves only for winter

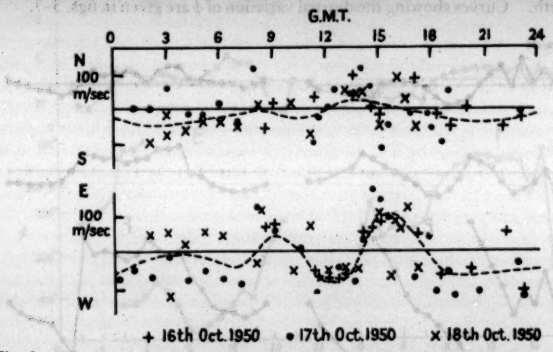


Fig. 2. Results of three days' observations by the method of closely spaced receivers, Cambridge, England. October 1950 (*after* Phillips).

Reproduced, by permission, from J. Atmos. Terr. Phys.

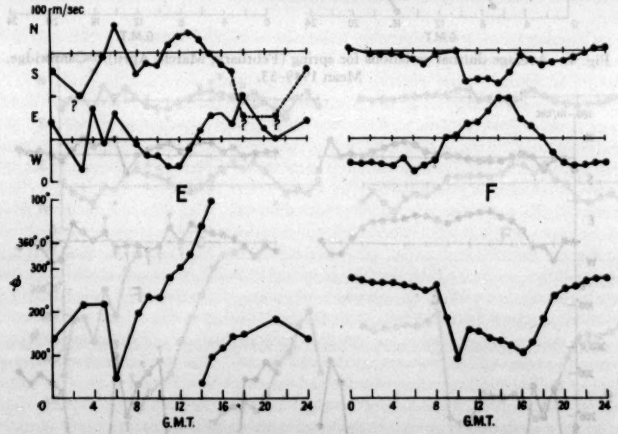


Fig. 3. Average diurnal variations for winter (November, December, January). Cambridge. Mean 1949–53.

(fig. 3) (average of November, December, January), spring (fig. 4) (average of February, March, April), summer (fig. 5) (average of May, June, July), and autumn (fig. 6) (average of August, September, October). Figure 7 shows the overall mean diurnal variations of the NS and EW components as deduced from all the observations.

From the curves showing the variations of the NS and EW components the average variation of the *direction* of movement was also found. We indicate the direction *towards which* the motion is directed by an angle ϕ measured clockwise from North. Curves showing the diurnal variation of ϕ are given in figs. 3–7.

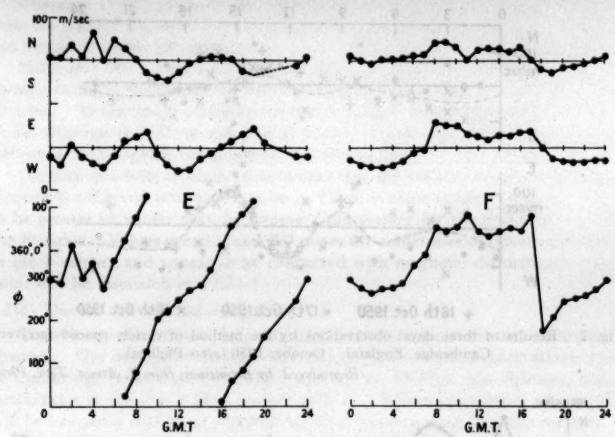


Fig. 4. Average diurnal variations for spring (February, March, April). Cambridge. Mean 1949–53.

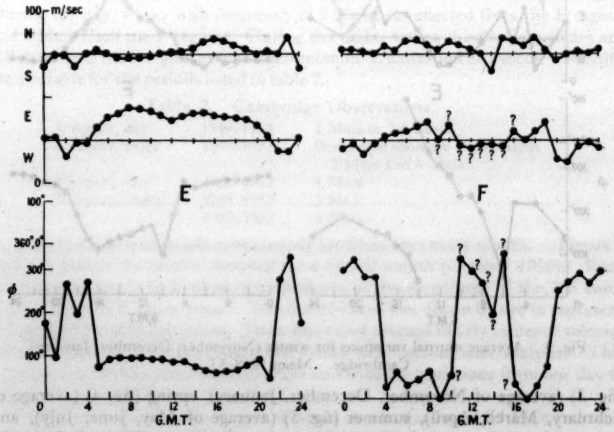


Fig. 5. Average diurnal variations for summer (May, June, July). Cambridge. Mean 1949–53.

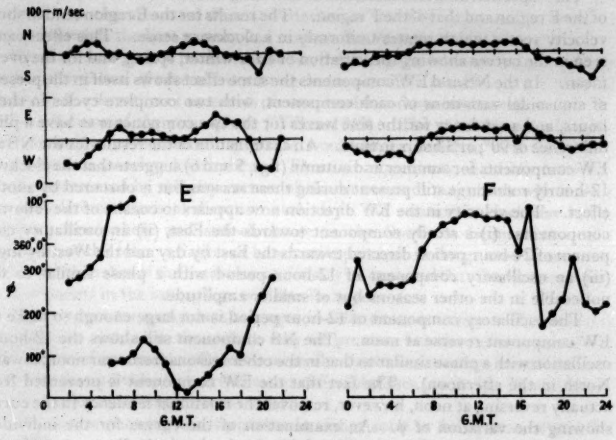


Fig. 6. Average diurnal variations for autumn (Aug., Sept., Oct.). Cambridge. Mean 1949–53.

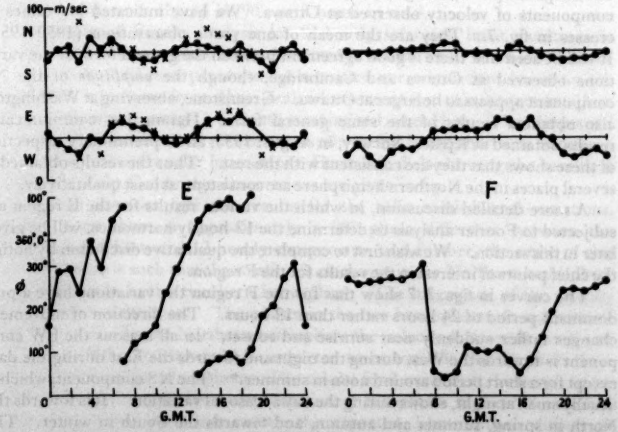


Fig. 7. Diurnal variations, overall mean. Cambridge 1949–53. The crosses are the mean values for one year's observations (1950–51) at Ottawa (*after* Chapman).

Figures 3–7 give the direction *towards which* the velocity is directed; the angle ϕ is measured clockwise from North.

An inspection of these results shows a striking difference between the behaviour of the E region and that of the F region. The results for the E region tend to show a velocity vector which rotates uniformly in a clockwise sense. This effect can be seen in the curves showing the variation of ϕ for winter, spring, and for the overall mean. In the NS and EW components the same effect shows itself in the presence of sinusoidal variations of each component, with two complete cycles in the 24 hours, and a tendency for the sine waves for the two components to have a phase difference of 90°, or 3 hours in time. An examination of the results for the NS and EW components for summer and autumn (figs. 5 and 6) suggests that the clockwise 12-hourly rotation is still present during these seasons but is obscured by another effect. The velocity in the EW direction now appears to consist of the following components: (i) a steady component towards the East, (ii) an oscillatory component of 24-hour period directed towards the East by day and the West by night, (iii) an oscillatory component of 12-hour period with a phase similar to that noticeable in the other seasons but of smaller amplitude.

The oscillatory component of 12-hour period is not large enough to make the EW component reverse at noon. The NS component still shows the 12-hourly oscillation with a phase similar to that in the other seasons (zero near noon, towards North in the afternoon). The fact that the EW component is prevented from actually reversing at noon, however, removes the rotational tendency in the curves showing the variation of ϕ. An examination of the curves for the individual months shows that the EW component is directed towards the East at noon for the months of May, June, July and August, and towards the West for the rest of the year.

Chapman (1953) has given the mean diurnal variations of the NS and EW components of velocity observed at Ottawa. We have indicated his values by crosses in fig. 7. They are the mean of one year's observations (1950–1951). It will be seen that there is good agreement between the general form of the variations observed at Ottawa and Cambridge, though the *amplitude* of the NS component appears to be larger at Ottawa. Greenstone, observing at Washington, also obtained results of the same general form. Harang has communicated results obtained at Kjeller, Norway, in August 1953, and a preliminary inspection of these shows that they are consistent with the rest. Thus the results obtained at several places in the Northern hemisphere are consistent, at least qualitatively.

A more detailed discussion, in which the various results for the E region are subjected to Fourier analysis to determine the 12-hourly harmonics, will be given later in this section. We wish first to complete the qualitative discussion by noting the chief points of interest in the results for the F region.

The curves in figs. 3–7 show that for the F region the variations have a predominant period of 24 hours rather than 12 hours. The direction of movement changes rather suddenly near sunrise and sunset. In all seasons the EW component is towards the West during the night and towards the East during the day, except for a short period around noon in summer.* The NS component, which is usually small at night, shows during the day a seasonal variation. It is towards the North in spring, summer and autumn, and towards the South in winter. The curves for the overall mean (fig. 7) show that for an average over the whole year

* The values for this period are based on relatively few observations and are unreliable. They are marked with ? on the curve. It should perhaps be emphasized that the daytime portions of all the curves for the F region are based on only one year's data (table 2).

the NS component is small, and shows little diurnal variation. There remains, however, the characteristic diurnal variation in the EW component, which is directed towards the East at noon and towards the West at midnight.

We shall now return to the results for the E region, in order to give a more detailed discussion of several interesting features. The presence of a 12-hourly rotation in the direction of motion is of considerable interest in connection with the tidal oscillations of the atmosphere (Wilkes 1949). The semi-diurnal solar and lunar variations of barometric pressure would be expected to have horizontal air movements associated with them. These air movements, in middle latitudes, can be represented by velocity vectors of almost constant magnitude which make two complete rotations of direction each day: the rotation is clockwise in the Northern hemisphere and anti-clockwise in the Southern hemisphere. The amplitude of the solar variation of barometric pressure at the ground is considerably larger than that of the lunar one, and it may be expected that the corresponding air movements in the ionosphere will also be larger.

Mean curves for many days averaged in solar time, as in figs. 3–7, will reveal any variation with solar time which may be present, but the variations in lunar time will tend to cancel. To determine the lunar variations it is necessary to return to the individual observations at different times (as in fig. 2), and group all the results according to lunar time.

The curves showing the variations in either solar or lunar time of the NS and EW components can be analysed by Fourier analysis to determine the amplitudes and phases of the 12-hourly harmonics. If the final result is to correspond to a vector of constant amplitude, with a uniform rotation of direction, the amplitudes of these harmonic components should be equal, and their phase difference should be 90°, or 3 hours in time.

The results of Fourier analysis for 12-hourly solar and lunar variations in the NS and EW components are summarized in tables 3 and 4. Some of these are taken from published papers, some represent our calculations based on our own results, and those from New Zealand represent our calculations based on results kindly sent to us by G. J. Burtt, working at Lower Hutt. The component curves for the New Zealand results are shown in fig. 8, together with harmonic dials for the 12-hourly solar harmonics. Harmonic dials showing the results of the analysis of the Cambridge observations are given in fig. 9.

We shall consider first the results for the solar semi-diurnal variations. In most cases there is approximately a 90° phase difference between the Northward and Eastward 12-hourly components, and further, the sense of this phase difference is such that the rotation is clockwise in the Northern hemisphere and anticlockwise in the Southern hemisphere. The magnitude of the velocity is of the order of 30 m/sec. The phases in the two hemispheres do not seem to agree very well. In the Northern hemisphere the velocity is towards the North at 0300 and 1500, and it would be expected to be towards the South at these times in the Southern hemisphere. The results for Lower Hutt, however, (if we take a mean which includes a consideration of the phases of the NS and EW components, which have not exactly the expected 3-hour phase difference) show it to be towards the South at 0500 and 1700. There is thus a discrepancy of 2 hours, or 60° in direction, in the results for the two hemispheres. It should be remembered that the results for Lower Hutt are based on only seven months' data.

Table 3. Solar 12-hourly Harmonics
(t = local solar hour angle)

Locality	Worker	Northward movement (m/sec)	Eastward movement (m/sec)	Extent of observations	
Ottawa 45° N, 75° W	Chapman (1953)	34 sin ($2t-10°$)	23 sin ($2t-87°$)	E region (day), 1950–51	
Cambridge 52° N, 1° E	Phillips (1952)	28 sin ($2t$)	28 sin ($2t-90°$)	E region (day), 1949–51	
	Spencer	Doubtful (<15 m/sec)		June 1952. One month's continuous observation on E region (day)	
	Cambridge results combined	40 sin ($2t+57°$)	45 sin ($2t-78°$)	Winter	E region (mainly day with some night-time sporadic-E), 1949–53
		30 sin ($2t-30°$)	39 sin ($2t-120°$)	Spring	
		6 sin ($2t-66°$)	20 sin ($2t-123°$)	Summer	
		16 sin ($2t+30°$)	6 sin ($2t-42°$)	Autumn	
		14 sin ($2t-12°$)	21 sin ($2t-93°$)	Mean	
Lower Hutt 41° S, 175° E	Burtt	31 sin ($2t+93°$)	29 sin ($2t+207°$)	E region (day), September 1952–March 1953	

Table 4. Lunar 12-hourly Harmonics
(τ = local lunar hour angle)

Locality	Worker	Northward movement (m/sec)	Eastward movement (m/sec)	Extent of observations
Ottawa 45° N, 75° W	Chapman (1953)	24 sin ($2\tau+3°$)	21 sin ($2\tau-106°$)	E region (mainly day), 1950–51
Washington 38° N, 77° W	Greenstone (private communication)	Not detectable (<10 m/sec)		E region (day), 1951–52
Cambridge 52° N, 1° E	Phillips (1952)	14 sin ($2\tau+4°$)	16 sin ($2\tau-81°$)	E region (day), 1949–51
	Spencer	Not detectable (<10 m/sec)		June 1952. One month's continuous observation E region (day)
	Spencer	Doubtful 14 sin ($2\tau+125°$)	Doubtful 12 sin ($2\tau-20°$)	September 1952. One month's continuous observation E region (day)

The general form of these results is in agreement with the predictions of the theory of atmospheric oscillations, but the phase of the rotation is not that expected theoretically. The maximum of the barometric pressure oscillation at the ground occurs at 1000. This corresponds to an air movement towards the West at ground level at this time; in the Northern hemisphere the movement would be towards the North 3 hours later, at 0100. It has been generally assumed, however, that the air movement near 100 km is out of phase with that at the ground (Wilkes 1949). This would correspond to a movement towards the North at 0700 in the Northern hemisphere and at 0100 in the Southern hemisphere.

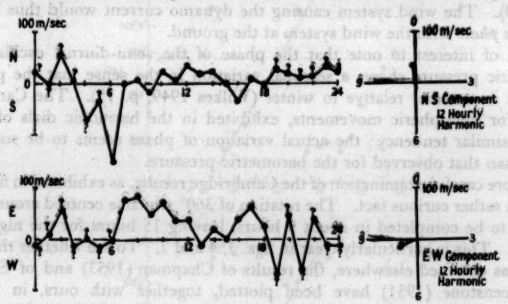


Fig. 8. Diurnal variations of NS and EW components of velocity at Lower Hutt, New Zealand (closely spaced receiver method). Mean for September 1952 to April 1953. E region only. 12-hourly harmonics were evaluated for the period 0600 to 1800: the results are shown on the harmonic dials to the right of the respective curves.

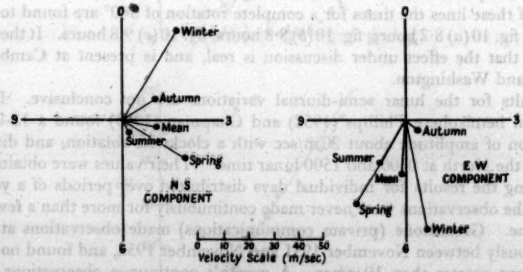


Fig. 9. Harmonic dials for the 12-hourly solar components for E region. Cambridge, 1949–52. Closely spaced receiver method.

Further evidence can be obtained from the semi-diurnal variations of the magnetic elements, which the dynamo theory associates with air movements in the ionosphere. For Northern latitudes the maximum Westward current density in the ionosphere occurs at 1100. If we assume that the currents are in phase with the dynamo e.m.f.s, this would require the air movement to be towards the North at this time.

The phase observed by the method of closely spaced receivers, towards the North at 0300, agrees with neither of the above results. The *amplitudes* of the semi-diurnal components would indicate a magnification, relative to the air velocity at ground level, of about 60 times.

The assumption mentioned above has recently been called in question by Fejer (1953). From a more complete treatment of the theory of the dynamo currents Fejer showed that the currents would be expected to lag about 1½ hours behind the associated dynamo e.m.f.s. The air movement, on the basis of the magnetic evidence, would then be towards the North, not at 1100 but at 1230 (or 0030). The wind system causing the dynamo current would thus be very nearly *in phase* with the wind system at the ground.

It is of interest to note that the phase of the semi-diurnal oscillation of barometric pressure shows a seasonal variation in the sense that the phase is retarded in summer relative to winter (Wilkes 1949, p. 15). The Cambridge results for ionospheric movements, exhibited in the harmonic dials of fig. 9, show a similar tendency; the actual variation of phase seems to be somewhat larger than that observed for the barometric pressure.

A more careful examination of the Cambridge results, as exhibited in figs. 3–7, reveals a rather curious fact. The rotation of 360° which is centred around noon appears to be completed in about 9 hours, leaving 15 hours for the night-time rotation. This is particularly clear in figs. 3, 4 and 7. To see whether this same result was obtained elsewhere, the results of Chapman (1953) and of Salzberg and Greenstone (1951) have been plotted, together with ours, in fig. 10. Figure 10(*a*) is a reproduction of the directional plot of fig. 7 (Cambridge). The points in fig. 10(*b*) were calculated from Chapman's results (Ottawa). Figure 10(*c*) is reproduced from a paper by Salzberg and Greenstone (1951, their fig. 6) (Washington). In each case a straight line has been drawn by eye through the points which lie symmetrically on either side of noon. From the slopes of these lines the times for a complete rotation of 360° are found to be as follows: fig. 10(*a*) 8·2 hours, fig. 10(*b*) 8·8 hours, fig. 10(*c*) 9·8 hours. It therefore appears that the effect under discussion is real, and is present at Cambridge, Ottawa and Washington.

Results for the lunar semi-diurnal variations are not conclusive. In the Northern hemisphere Phillips (1952) and Chapman (1953) found a 12-hourly oscillation of amplitude about 20 m/sec with a clockwise rotation, and directed towards the North at 0300 and 1500 lunar time. Their values were obtained by combining the results for individual days distributed over periods of a year or more; the observations were never made continuously for more than a few days at a time. Greenstone (private communications) made observations at noon continuously between November 1951 and November 1952, and found no lunar component greater than 10 m/sec. A month's continuous observations at all hours of the day at Cambridge in June 1952 gave a similarly negative result. A second month's observations in September 1953 gave a vector rotating clockwise and directed towards the North at 1100 and 2300 lunar time; the magnitude was about 13 m/sec, and the statistical significance of the result was not very high. No results for lunar variations are available from the Southern hemisphere.

It should be pointed out that determinations which are based on scattered observations over a long period are open to doubt, in that the apparent lunar component could be produced by a seasonal variation in the solar component.

Millman (1952) has published his results, obtained by reflection of waves of frequency 150 kc/s from the E region, in the form of diurnal curves for several months in 1951 and 1952. We have combined these curves to give a mean diurnal variation, and used Fourier analysis to search for evidence of a rotation with a 12-hour period, but without success. Millman stated that some of the curves seemed to show semi-diurnal variations, but did not give the results of a Fourier analysis.

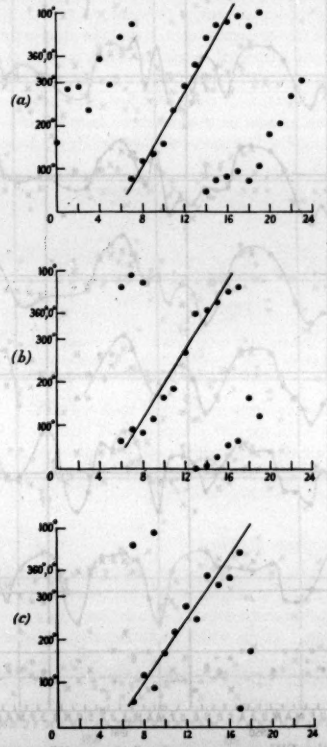


Fig. 10. Comparison of diurnal variation of direction of movement in E region at different places: (*a*) Cambridge, 1949–53. (*b*) Ottawa, 1950–51. (*c*) Washington, March, April and May 1950.

Bowhill (1953), observing on 70 kc/s in the summer months of 1952, found no evidence of a significant diurnal variation. The mean movement was towards the NE for the period of observation, which was 2000–0400.

We shall now return to the Cambridge results in order to exhibit the diurnal and seasonal changes in a slightly different way. In figs. 11 and 12 we have selected a particular hour of the day and have plotted the values for that hour

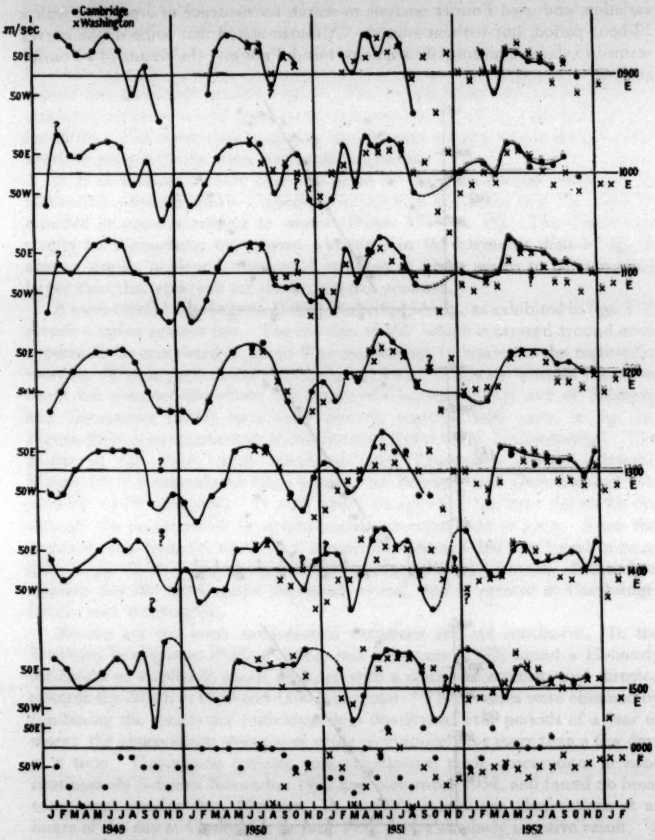

• Cambridge
× Washington
m/sec
50E
50W
0900 E
1000 E
1100 E
1200 E
1300 E
1400 E
1500 E
0000 F
J F M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O N D J F
1949
1950
1951
1952

Fig. 11. Variation of the EW component of velocity at different hours of the day, as observed at Cambridge, 1949–53. Observations made at the same local times at Washington are indicated by crosses.

successively, one point per month, throughout the four years of observation. The curves for 0900 to 1500 are for the E region only. The bottom curve is for the F region at 2400, and is characteristic of the behaviour of this region for the greater part of the night.

Consider first fig. 11, which refers to the EW component. The curve for 1200 (E region) shows very clearly a seasonal change in the EW component. The movement at noon is towards the East in summer and towards the West in winter.

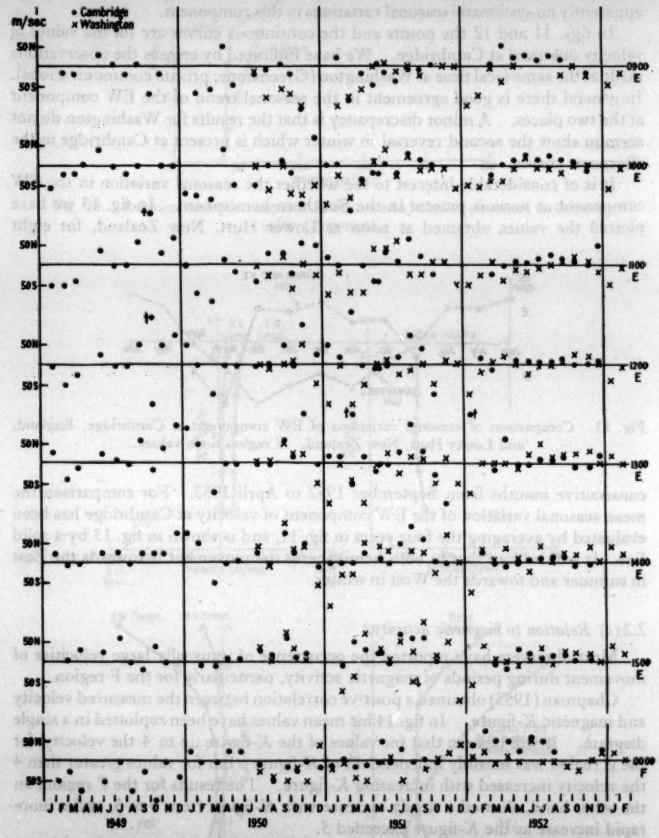

1 m/sec
Cambridge
Washington
50N
50S
0900 E
1000 E
1100 E
1200 E
1300 E
1400 E
1500 E
0000 F
J F M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O N D J F
1949
1950
1951
1952

Fig. 12. Variation of the NS component of velocity at different hours of the day, as observed at Cambridge, 1949–53. Observations made at the same local times at Washington are indicated by crosses.

We have already seen how the mean movement towards the East in summer is great enough to override the 12-hourly oscillation, which would otherwise produce a movement towards the West near noon.) For the morning and afternoon hours

the curves become more complicated, and in the afternoon curves there is a suggestion of a second reversal of the EW component in the middle of winter. The curve for the E region at 2400 shows no significant seasonal variations.

Figure 12 is plotted in the same way for the NS component. There are apparently no systematic seasonal variations in this component.

In figs. 11 and 12 the points and the continuous curves are for the values of velocity observed at Cambridge. We have indicated by crosses the observations made at the same local time at Washington (Greenstone, private communications). In general there is good agreement in the seasonal trend of the EW component at the two places. A minor discrepancy is that the results for Washington do not seem to show the second reversal in winter which is present at Cambridge in the afternoon curves.

It is of considerable interest to see whether the seasonal variation in the EW component at noon is present in the Southern hemisphere. In fig. 13 we have plotted the values obtained at noon at Lower Hutt, New Zealand, for eight

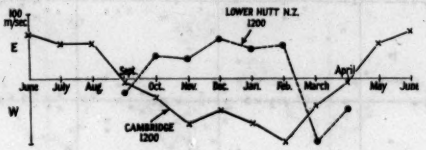


Fig. 13. Comparison of seasonal variations of EW component at Cambridge, England, and Lower Hutt, New Zealand. E region noon values.

consecutive months from September 1952 to April 1953. For comparison the mean seasonal variation of the EW component of velocity at Cambridge has been evaluated by averaging the four years in fig. 11, and is shown in fig. 13 by a solid line. It will be seen that in both hemispheres the movement is towards the East in summer and towards the West in winter.

2.2 (*c*) *Relation to magnetic activity.*

Most observers have reported the occurrence of unusually large velocities of movement during periods of magnetic activity, particularly for the F region.

Chapman (1953) obtained a positive correlation between the measured velocity and magnetic *K*-figure. In fig. 14 his mean values have been replotted in a single diagram. It will be seen that for values of the *K*-figure up to 4 the velocity for the E region was sensibly independent of *K*-figure; but for values greater than 4 the velocity increased with increasing *K*-figure. The results for the F region, on the other hand, showed a steady increase with *K*-figure from 0 to 5, and a more rapid increase as the *K*-figure exceeded 5.

Results obtained at Cambridge between 1949 and 1953 have been analysed to determine the dependence of the NS and EW components of the velocity and the total velocity on the magnetic *K*-figure (fig. 15). For the E region both components and the total velocity are independent of *K*-figure for values between 1 and 5; results for other *K*-figures are too few to be significant. (High velocities of the order of 500 m/sec were, however, recorded in the E region on one occasion when

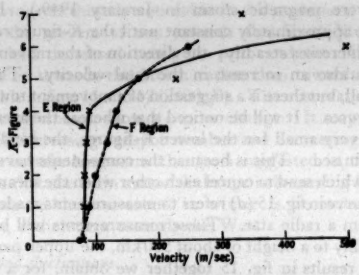


Fig. 14. Relation between magnetic *K*-figure and velocity observed at Ottawa (*after* Chapman).

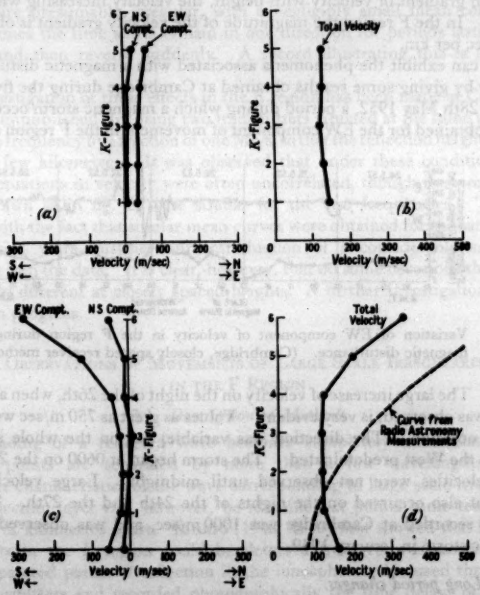


Fig. 15. Relation between magnetic *K*-figure and velocity observed at Cambridge: (*a*) E region EW and NS component, (*b*) E region total velocity, (*c*) F region EW and NS component, (*d*) F region total velocity.

there was a severe magnetic storm in January 1949.) For the F region velocities remain approximately constant until the *K*-figure reaches 5, when the EW component increases steadily; the direction of the movement is towards the West. There is also an increase in the total velocity. The NS component remains very small, but there is a suggestion of a movement towards the South for the highest *K*-figures. It will be noticed that whereas the average values of both components are very small for the lower *K*-figures, the total velocity is always greater than 100 m/sec. This is because the components have both positive and negative values, which tend to cancel each other when the mean is taken.

The broken curve in fig. 15 (*d*) refers to measurements made by Hewish (1952) on the waves from a radio star. These measurements will be discussed in § 4. They refer probably to a height of about 500 km, the upper part of the F region. If we take all the results in fig. 15 together we obtain, for a *K*-figure of 5, the following values of velocity at different heights.

Height (km)	100	300	500
Velocity (m/sec)	100	200	500

These results suggest, then, that at times of magnetic activity there is a pronounced gradient of velocity with height, the velocity increasing with increasing height. In the F region the magnitude of the velocity gradient is of the order of 1·5 m/sec per km.

We can exhibit the phenomena associated with a magnetic disturbance more directly by giving some results obtained at Cambridge during the five days from 24th to 28th May 1952, a period during which a magnetic storm occurred. The results obtained for the EW component of movement in the F region are shown in

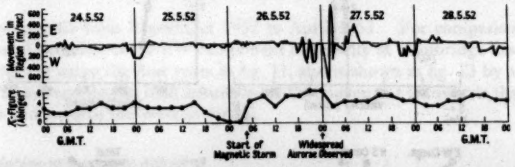


Fig. 16. Variation of EW component of velocity in the F region during a period of magnetic disturbance. (Cambridge, closely spaced receiver method.)

fig. 16. The large increase of velocity on the night of the 26th, when a widespread aurora was observed, is very evident. Values as great as 750 m/sec were recorded on this occasion. The direction was variable, but on the whole a movement towards the West predominated. The storm began at 0600 on the 26th, but the large velocities were not observed until midnight. Large velocities around midnight also occurred on the nights of the 24th and the 27th. The highest velocity recorded at Cambridge was 1000 m/sec, and was observed during the magnetic storm in January 1949.

2.2(*d*) *Long period changes.*

Observations at Cambridge have now been made for four years. When these results are plotted consecutively, as in figs. 11 and 12, certain long period changes become evident.

The most obvious of these is noticeable in the movement in the F region at night. In 1949 the movement was towards the South-West. By 1952 the EW component had decreased nearly to zero, and the NS component had reversed so as to be towards the North rather than towards the South. This change is noticeable in figs. 11 and 12 (2400, F). The observations at Washington, which are shown by crosses, show a similar change. It may be noted that this period was one of decreasing sunspot number and decreasing magnetic activity.

A rather less certain long-term change is suggested in the other curves of fig. 11. It seems that the main seasonal change is becoming less definite. This is happening because the movement in winter is becoming more and more towards the East, rather than towards the West.

2.2(*e*) *Short period fluctuations.*

In the preceding discussion we began by taking smoothed hourly values of the velocity, and variations of shorter period were thereby removed. More rapid fluctuations do, however, occur, and have been studied at Cambridge.

On some occasions rapid fading is observed, but without any consistency of the magnitudes or senses of the time shifts. On these occasions the ionospheric velocities are presumably irregular, with a negligibly small mean component. At other times the time shifts remain in one direction for periods lasting a few minutes, and then reverse suddenly. A record illustrating this is shown in fig. 1 (*b*).

An investigation of these effects in the E region has been made at Cambridge (Spencer, unpublished) by using two transmitters situated at the same point, but differing in frequency by a fraction of one Mc/s, so that the reflection heights differed by only a few kilometres. It was observed that under these conditions, short period fluctuations in velocity were often uncorrelated, though the *mean* diurnal curves, drawn as in fig. 2, were similar for the two frequencies. This fact, together with the fact that similar mean curves were obtained for the same month in successive years, justifies our introduction of a considerable amount of smoothing into the data. It is clear, however, that on some occasions the velocity is markedly different at closely spaced heights. A further investigation of these points is in progress.

§ 3. Observations of Movements of Large Scale Irregularities in the F Region

3.1. *Description of Methods*

The fading of a reflected wave becomes completely uncorrelated if the receiving points are separated by more than a few wavelengths. Other measurable characteristics of the wave corresponding to large scale irregularities in the reflecting layer can, however, be identified at points some tens or even hundreds of kilometres apart. Munro (1948, 1950) first showed that horizontal movements in the F region could be detected from records associated with widely separated points of reflection in the ionosphere. He used three distant pulse transmitters and recorded photographically the time delays of the three echoes returned from the F region to a single receiving point. In this way the virtual height of the reflecting layer was measured at three different points. The following characteristics were found suitable for the detection of horizontal

movements: (*a*) virtual height peaks, (*b*) virtual height dips, (*c*) cross-overs of the ordinary and extraordinary waves, (*d*) an effect in which the echo was observed to split into three components with different time delays.

If the waves from the three transmitters were found to exhibit any of these phenomena of similar form, but at different times, then a measurement of the time intervals between their occurrences enabled the velocity of horizontal movement to be determined.

Munro has reported the occurrence of a particular type of travelling disturbance in the F region, the form of which could be determined in detail from a study of (P', f) curves. At a fixed height the electron density was found first to decrease, then to increase, and finally to return to normal. The whole cycle was completed in about half an hour. The changes at other heights were similar except that there was a variation of phase with height, in the sense that the changes took place later at the lower height. The same disturbances have been identified (with time shifts) on (P', f) records obtained at points separated by as much as 900 km, so they must be capable of travelling this distance without sensible change of form. Munro considers that the changes in the F region listed under (*a*), (*b*), (*c*), (*d*) above are due to smaller travelling disturbances of the same type. This view would appear to imply that the velocity of some type of compressional wave is involved rather than a bodily movement of the medium.

Bramley (1953) used a different method to detect horizontal movements in the F layer. He used direction-finding techniques to study the 'tilt' of the layer, and showed that the variations of the tilt, as observed at two points separated by 27 km, were often displaced in time.

Bramley also developed a method in which simultaneous measurements of the bearing and elevation of the signals from a distant transmitter were made at a single station. He showed that it was possible to use the measurements to detect the passage of 'ripples' in the F region, and to deduce (i) their speed of travel, (ii) their direction of travel, (iii) their horizontal 'wavelength', (iv) their vertical amplitude. These ripples are almost certainly the same as the travelling disturbances observed by Munro.

Movements have been deduced from observations at widely spaced points by Beynon and Thomas (private communication). They used a network of three stations at Slough, Swansea and Bangor. At each point complete normal incidence (P', f) curves were recorded at frequent intervals. Irregularities in the F region, such as changes of height or critical frequency, were noted at each station and examined for the presence of time shifts. The distances involved in this experiment were as follows: Slough–Swansea, 232 km; Swansea–Bangor, 180 km; Bangor–Slough, 308 km. In later experiments a fourth station at Inverness, 650 km north of Swansea, was used. Corresponding disturbances could sometimes be identified at all four stations.

The methods which use widely spaced reflection points have given results for the F region, but have not so far been used for the normal E region. This is because there do not seem to be irregularities of sufficiently large scale in this region. An exception to this statement arises when sporadic-E ionization is present in the E region; this occurs in patches with a size of the order of a few hundred kilometres. It is then possible to detect apparent horizontal movements by observing the incidence of the sporadic-E ionization at widely spaced points. We shall discuss this method in § 6.

3.2 *Results*

The most detailed published results are those of Munro (1950), working at Sydney, Australia (virtual height measurements with three distant transmitters). He observed mainly in the daytime (0900–1600), and found velocities in the F region which were usually between 80 and 200 m/sec. The direction of movement was approximately the same throughout any one day, but had a well defined seasonal variation. Figure 17 shows the mean direction of movement for each

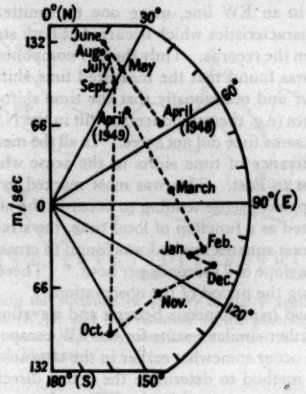


Fig. 17. Seasonal variation of velocity in the F region at Sydney (*after* Munro).
Reproduced by courtesy of the Royal Society.

month from April 1948 to April 1949. It will be seen that there was a component towards the East at all seasons, but the NS component was oppositely directed in summer and winter. If the summer and winter months are averaged, the following mean directions are obtained :

Winter : mean direction 20° E of N
Summer : mean direction 110° E of N

More extended measurements (Munro 1953) showed a small diurnal variation of direction, most marked in the month of June. For this month observations were available for three successive years (1950–52). If the directions for the three years are averaged, it is found that the direction changed from 50° E of N at 0930 to 15° E of N at 1430. A variation in the same sense was evident in June for each year taken alone. Some observations made at night showed a consistent movement towards the West, in contrast to the movement towards the East observed in the daytime at all seasons. The actual times of reversal of the EW component were not stated. If, however, the diurnal change just described is assumed to continue at the same rate, it will be seen that the EW component will have decreased to zero at about 1630. For comparison with other results we may therefore provisionally assume that the EW component reverses at this time.

Price (1953), working at Nedlands, Western Australia, made measurements similar to those of Munro, and observed the same seasonal change of direction. In collaboration with Munro, he showed that the velocities measured were sometimes the same at points separated by 2000 miles.

Bramley (1953), working at Slough, England, has given some results obtained by the two different methods described above. The results related to the period November 1949 to July 1950, and were confined mainly to times between 1000 and 1700. In the first method the tilt of the F layer was observed at two points separated by 27 km in an EW line, using one transmitter. The instances of clearly identifiable characteristics which occurred at both stations were examined for time shifts between the records. Only the EW component of movement could be determined. It was found that the measured time shifts at any time of day showed a wide scatter and occasionally that the time shifts associated with two different characteristics (e.g. the components of tilt in the NS and EW directions) observed at about the same time did not agree. In all the measurements, however, there was a preponderance of time shifts in the sense which would indicate a movement from West to East. This was most marked in the morning, with a change-over to the opposite sense tending to occur in the afternoon. When the time shifts were plotted as a function of local time, the straight line fitted to the observations by the least squares method was found to cross the axis of zero time shift at 1520 and had a slope of 23 seconds per hour.* There was no evidence of a seasonal change during the period of the observations.

The second method (simultaneous bearing and elevation measurements at a single station) gave rather similar results for the EW component, except that the reversal was found to occur somewhat earlier in the afternoon, at about 1400. It was possible by this method to determine the actual direction of travel, and it appeared that the afternoon reversal in the EW component was really due to a clockwise rotation in direction (East to West via South) rather than a change in the magnitude of the velocity. Bramley has not reported the seasonal change in the NS component which is observed by Munro and Price.

If we compare the results of Munro, Price and Bramley with those obtained by the closely spaced receiver method for the F region (§ 1, figs. 3–7), we see that there is substantial agreement for the direction of the movement. Thus the EW component as determined by the closely spaced receiver method is towards the West by night and towards the East by day at all seasons (figs. 3–7). There is a possible exception of some doubtful summer results in fig. 5. The rotational effect, however, does not show itself, and the afternoon reversal of the EW component seems to take place rather later in the day. It should be mentioned here that the daytime results for the F region from the closely spaced receiver method are based on one year's observations only (table 2), and so are less reliable than the rest of the observations. In particular, the precise times of the diurnal reversal of the EW component cannot be regarded as well established.

The NS component determined by the closely spaced receiver method shows in the daytime a seasonal reversal, being towards the North in summer and towards the South in winter. This would agree with the "South in summer, North in winter" observed by Munro and Price in the Southern hemisphere. The effect does seem to be a little different in that the reversal of the NS component is confined to the three winter months (fig. 3), whereas their reversals take place symmetrically at the equinoxes (fig. 17).

* In this experiment a time shift of 1 minute indicated a speed of movement of 210 m/sec.

The magnitudes of the velocities observed by the above methods appear to be somewhat larger than for the closely spaced receiver method (§ 2.2 (*a*)).

Munro, Price and Bramley have not reported any dependence of velocity on magnetic activity. Beynon and Thomas (private communication), however, found that the velocity in the F region as determined from their network of (P',f) stations was closely related to the degree of magnetic disturbance. Figure 18 shows the mean velocities obtained from a study of the variations in

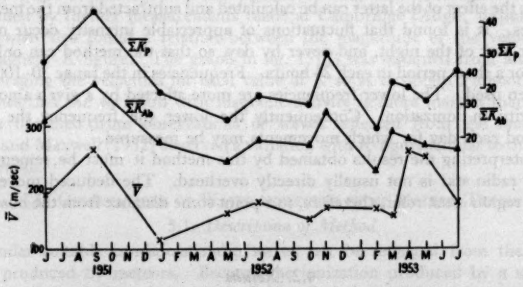


Fig. 18. Curves showing the variation of velocity in the F region, the variation of the Abinger *K*-figure (K_{Ab}), and the variation of the planetary *K*-figure (K_p). The *K*-figures plotted are average daily sums for the days of the ionospheric observations (*after* Beynon and Thomas).

the normal incidence critical frequency of the F region. Each point in the diagram represents the mean of 10 to 15 measurements made during the daytime over periods of three consecutive days. In the same figure are shown the variation in the Abinger *K*-figure K_{Ab} and in the planetary magnetic disturbance index K_p published by Bartels. The *K*-figures plotted are average daily sums for the days of the ionospheric observations. Further confirmation of the correlation between velocity and magnetic disturbance was obtained from a study of detailed variations over periods of a few hours. A number of cases were observed in which a short period change in magnetic activity was simultaneous with a change in the magnitude of the velocity; more often, the velocity change followed an hour or two later. A study of these changes showed that at times of increased magnetic activity the movement in the F region was frequently towards the South.

Beynon and Thomas also reported a marked gradient of velocity with height in the F region, of the order of 1 m/sec per km; the velocity increased with increasing height. At the maximum of the F layer the mean velocity they obtained was 190 m/sec.

§ 4. Observations of Waves from Radio Stars

4.1. *Description of Method*

The intensity and apparent position of a radio star when observed at metre wavelengths are often found to fluctuate in an irregular manner. It has been shown that the fluctuations are caused by the passage of the waves through

regions of irregular ionization in the upper part of the F region (Smith 1950, Little and Lovell 1950, Ryle and Hewish 1950, Little and Maxwell 1951, Little 1951, Hewish 1951). When observations are carried out with similar receiving equipments separated by a few kilometres it is possible to detect differences in the times of occurrence of the individual fluctuations recorded at the various stations. These time displacements are the result of the drift of a diffraction pattern across the ground. This is attributed partly to the movement of irregularities in the F region and partly to the rotation of the Earth relative to the star; the effect of the latter can be calculated and subtracted from the measured velocities. It is found that fluctuations of appreciable intensity occur only in the later part of the night, and never by day, so that the method can only give results for a short period in each 24 hours. Frequencies in the range 30–100 Mc/s have been used. The lower frequencies are more affected by a given amount of irregularity in ionization. Consequently the lower the frequency the larger the period each day for which movements may be measured.

In interpreting the results obtained by this method it must be remembered that the radio star is not usually directly overhead. The deduced movements of the F region often refer, therefore, to a point some distance from the observing station.

4.2. *Results*

Results of observations of waves from radio stars have been published by Maxwell and Little (1952), working at Manchester, and by Hewish (1952), working at Cambridge. Further results have been obtained at Cambridge by Adgie and Newstead (unpublished).

During May and June 1951 Maxwell and Little found a movement mainly towards the West. Their observations, made between 1800 and 0100, indicated a mean velocity of 70 m/sec in a direction 15° N of W. They referred to a part of the ionosphere at a mean latitude of 60° N, i.e. some 800 km north of the receiving equipments. During September and October 1951 the radio stars were near the zenith and the measurements referred to conditions at a latitude of 53° N. The mean velocity was 120 m/sec, predominantly towards the West, but with a small component towards the South (mean direction 15° S of W). It is not known whether the reversal in the NS component of velocity is associated with change of season or of latitude.

Results given by Hewish appear at first sight to differ from those of Maxwell and Little, since Hewish gives a histogram (his fig. 12) which shows a roughly equal number of velocities directed towards the East and towards the West, with, if anything, a preponderance of velocities towards the East. These results are for the period January to April 1951 and were made between 0000 and 0500 on most nights. Further consideration shows that the discrepancy is not as serious as appears at first sight. In fact, for velocities of the order of 100 m/sec (the order of magnitude obtained by Maxwell and Little) the histogram given by Hewish shows a preponderance of velocities towards the West. The greater velocities (200–500 m/sec) are, as shown by Hewish, associated with magnetic activity, and it is these high velocities which show the predominant movement to be towards the East. It is interesting to note that Maxwell and Little observed only before midnight and Hewish only after midnight.

It appears that the normal direction of movement of the upper part of the F region at night is towards the West, but that during periods of magnetic disturbance the movement is towards the East after midnight. This is confirmed by observations made by Hewish on six successive nights of severe magnetic disturbance from 2nd–9th April 1951. On each night the direction of movement was towards the West at 1900 and towards the East at 0100, and the change took place near midnight. This tendency for a reversal of the EW component to take place at midnight during periods of magnetic disturbance has been confirmed by further measurements made at Cambridge (Adgie, unpublished).

Hewish also found a relation between the magnitude of the velocity and the magnetic *K*-figure. The graph in fig. 15(*d*) was obtained from his fig. 13, by averaging the velocities for each value of K so as to obtain a smooth curve. It shows that the variation with magnetic activity is more marked than for the results obtained from observations of waves reflected from the ionosphere. Little and Maxwell (1952) observed a variation with *K*-figure similar to that found by Hewish.

§ 5. Observations of Waves reflected from Meteor Trails

5.1. *Description of Method*

Under suitable conditions radio waves can be reflected from the ionized trails produced by meteors. Because the ionization produced by a meteor is distributed along a line, the strength of the reflected signal depends greatly on the orientation of the column. Normally a reflection is only obtained when the trail is perpendicular to a ray from the observing station. Immediately upon the passage of a meteor a burst of reflected signal is received lasting usually for about a second. Occasionally reflections of longer duration are obtained; some may last for a few minutes. The reflection points vary in height from 80 to 110 km on different occasions.

If the ionized trail drifts after it is formed, its distance from the observing station will change with time, and the radial component of velocity can be measured. If the drift is due to a uniform horizontal movement of the atmosphere, then two trails in different directions from the station will have different radial components of velocity, and measurements of the two positions and radial velocities are sufficient to determine the magnitude and direction of the horizontal movement. The assumption of a uniform horizontal atmospheric movement can be verified if a large number of meteor trails can be observed in a small interval of time in directions all around the observing station. Then, if the horizontal component of the radial velocity is plotted against the azimuthal angle, a sine curve should be obtained.

Various methods have been used to determine the radial velocities and the positions of the trails in space. Manning, Villard and Peterson (1950) used a continuous-wave Doppler method to measure the velocity, together with a special direction finder to locate the trails. Greenhow (1952) used a pulse method, and was able to measure the delay time of an echo with sufficient accuracy to determine the change in range during the lifetime of the echo. He could, however, measure velocities only along a North–South line. Robertson, Liddy and Elford (1953) devised a method which was a combination of the pulse and continuous wave methods, by which the radial velocity and location of a trail could both be measured.

5.2. *Results*

Manning, Villard and Peterson (1950), working at Stanford, California (37° N, 123° W), made observations on about 18 occasions during the summer of 1949. Observations were made before sunrise between the hours 0300 and 0600, because there would be a high rate of occurrence of meteors at that time and a low level of interference. The horizontal component of the observed radial velocity was plotted against its direction, and the sine curve which best fitted the results was determined statistically. From the amplitude and position of this sine curve the magnitude and direction of the horizontal velocity could be determined. It was found that the most frequent directions of movement were towards the North-East and North, with velocities of the order of 50 m/sec. The heights of the trails were between 90 and 110 km. It was found that the deviations of the individual points from the best-fitting sine curve were considerable, and it was deduced that there were irregular components of velocity at least as large as the mean.

Greenhow (1952), working at Manchester, made observations on the NS component of velocity during the nights of 12th–13th and 13th–14th December 1952, when the Geminid meteor shower was active. Between 2300 and 2400 the direction of movement was towards the South. From 2400 to 0200 no well-defined NS component of velocity could be determined. This result indicated either that the movement was irregular or that it was predominantly in the EW direction. No observations were made between 0200 and 0320, but after 0320 most of the echoes indicated a movement towards the North. The NS component of velocity appeared, therefore, to have reversed at about 0100, being towards the South before that time and towards the North afterwards. The northward velocity after 0320 is in agreement with the results of Manning *et al.*

We can compare these results with those obtained at Cambridge by the closely spaced receiver method for the E region (§ 2.2(*b*)). We saw that, on the average, the predominant movement at this height could be represented by a velocity vector of constant magnitude (~30 m/sec) which made two complete rotations every 24 hours, the phase being such that the movement was towards the North at 0300 and 1500. We have illustrated this in idealized form in the NS and EW component curves of fig. 19. (These may be compared with the experimental results of figs. 3–7.) The times at which the meteor observations were made are marked in fig. 19. It will be seen at once that the results obtained from meteors in the Northern hemisphere are in agreement with the results of the closely spaced receiver method. There is one minor discrepancy: the results of Greenhow suggested that the NS component reversed at about 0100, whereas the curves of fig. 7 suggest that it should reverse at midnight. It should be pointed out that the closely spaced receiver method normally only gives results for the E regions in the daytime. The night-time parts of the curves in fig. 19 are, therefore, essentially extrapolations, though they are confirmed by a few isolated measurements on sporadic-E ionization made during the night.

We shall now describe the results obtained by Elford and Robertson (1953) at Adelaide (35° S, 139° E). The observations were made during 13th–30th October, 12th–30th November and 2nd–20th December 1952, and were

essentially continuous between these dates. In general, the echo rate was too low to permit a good determination of the horizontal velocity at any particular hour. It was, however, possible to determine the average *diurnal variation* of

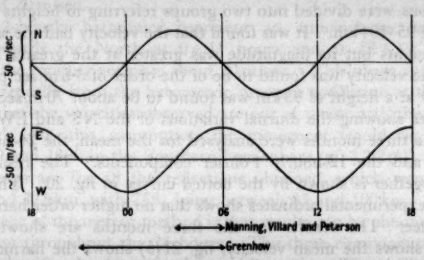


Fig. 19. Idealized diurnal variations as observed at Cambridge by the closely spaced receiver method. The arrows indicate the periods during which measurements were made by workers using the meteor method in the Northern hemisphere.

the velocity for each of the three periods of observation by combining the results for all meteors observed during a given hour on the various days. The observations combined in this way were found to show the expected sinusoidal variation with azimuth, with, however, a considerable scatter. This scatter could be attributed to three causes: (i) day to day variations in the velocity at the same hour, (ii) random variations of the velocity, (iii) variation of the velocity with height.

The mean velocity was shown to be horizontal to an accuracy of at least 5°. When the diurnal variations of the NS and EW components are plotted, a remarkable general similarity is observed between them and the corresponding results obtained by the closely spaced receiver method. In order to bring out this similarity we have reproduced in fig. 20 Elford and Robertson's curve for

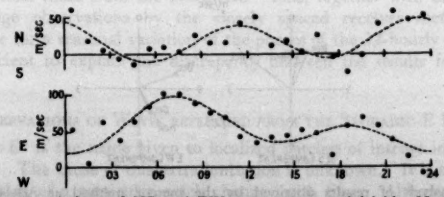


Fig. 20. Diurnal variations of NS and EW components at Adelaide, November 1952. Meteor method (*after* Elford and Robertson).

the period 12th–30th November. This curve for a summer month may be compared with the summer results at Cambridge (fig. 5) and also with the New Zealand results for summer (fig. 8). In all cases there is a mean movement

towards the East, with evidence of components of 12- and 24-hour periods superimposed.

The curve of fig. 20 was drawn for all the observations, irrespective of the height to which they referred. To investigate a possible change with height the observations were divided into two groups referring to heights in the ranges 85–94 km and 95–104 km. It was found that the velocity had the same direction at the two heights but its magnitude was greater at the greater height. The gradient of the velocity was found to be of the order of 3·6 m/sec per km. The mean velocity at a height of 95 km was found to be about 70 m/sec.

The curves showing the diurnal variations of the NS and EW components for each of the three months were analysed for the mean, the 24-hourly Fourier components, and the 12-hourly Fourier components. The result of adding these three together is shown by the dotted curves in fig. 20. The fit of these curves to the experimental ordinates shows that no higher order harmonics are of any importance. The results for the three months are shown in fig. 21. Figure 21 (*a*) shows the mean velocity, fig. 21 (*b*) shows the harmonic dials for

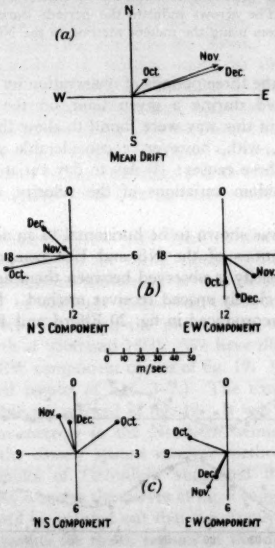


Fig. 21. Analysis of results obtained by the meteor method at Adelaide, October, November and December 1952: (*a*) mean drift, (*b*) 24-hourly harmonic dials, (*c*) 12-hourly harmonic dials. (*After* Elford and Robertson.)

the 24-hourly harmonics, and fig. 21 (*c*) shows the harmonic dials for the 12-hourly harmonics. The results may be summarized as follows:

(*a*) In November and December (summer) the mean velocity is towards a direction slightly North of East, and has a magnitude of about 50 m/sec, while

in October it is small. Comparison with figs. 3–7 shows that this is consistent with the results from closely spaced receivers (§ 2.2(*b*)).

(*b*) The 24-hourly harmonics represent an anticlockwise rotation, with movement towards the North at 18 ± 2 hours. The magnitude of the rotating vector is about 20 m/sec.

(*c*) The 12-hourly harmonics represent an anticlockwise rotation, with movement towards the North at $0 \pm 1\frac{1}{2}$ hours and at $12 \pm 1\frac{1}{2}$ hours. The magnitude of the rotating vector is about 20 m/sec. This phase agrees with that expected theoretically from the barometric pressure oscillation at the ground, if one assumes that the velocities observed are out of phase with the air velocities at the ground. On this assumption the movement would be towards the North at 0100 and 1300 in the Southern hemisphere (cf. § 2.2(*b*)).

These results are for all the reflections obtained, which were for heights between 75 and 115 km, with 75% lying between 85 and 100 km.

An advantage of the meteor method is that results can be obtained for heights in the E region throughout the 24 hours. The three components (*a*), (*b*) and (*c*) above can therefore be determined unambiguously. The results from closely spaced receivers, on the other hand, give results for the E region only during the daytime on the frequencies normally used, apart from occasional night-time sporadic-E. Consequently the 12-hourly harmonics may be determined with some accuracy, but the mean velocity and the 24-hourly harmonics cannot be accurately determined. The only reliable comparison which can be made is, therefore, between the amplitudes and phases of the 12-hourly harmonics as determined by the two methods. When the results for the meteor method (fig. 21) are compared with those obtained at Lower Hutt, New Zealand, by the closely spaced receiver method (fig. 8) it is found that there is good agreement. At Lower Hutt the analysis gave a velocity vector of magnitude 30 m/sec, rotating anticlockwise, and directed towards the North at about 1100 and 2300 hours local time.

We showed in § 2.2(*b*) that the results for Lower Hutt were inconsistent with the results obtained by the same method in the Northern hemisphere in that there was a difference of 2 hours in the phase of the 12-hourly harmonics. It should be pointed out that the results from the Southern hemisphere are based on fewer observations than those from the Northern. This, together with the fact that the Cambridge observations by the closely spaced receiver method show some evidence for a seasonal variation of the phases of the 12-hourly harmonics, may be sufficient to explain the discrepancy between the results for the two hemispheres.

§ 6. Observations of Waves reflected from the Sporadic-E Layer

'Sporadic-E' is the name given to localized patches of intense ionization in the E region. The cause of this extra ionization is unknown. It is often found that the position of a particular patch changes with time. Although it is not possible at present to say whether this effect is due to a motion of the cause or to a drift of the patch produced by air movements, the phenomenon is clearly of interest in the present context.

Movements of patches of sporadic-E ionization have been detected in various ways. Munro (1950) observed the times of appearance and disappearance of sporadic-E reflections from three distant transmitters (§ 3.1) and found that they

sometimes showed delays of the type expected if a patch of extra ionization had moved horizontally. In America radio amateurs observed when signals were propagated to long distances on a frequency of 50 Mc/s. If two distant stations could transmit to each other on this frequency, it was known that sporadic-E ionization must exist near the mid-point of the line joining them. The number of stations was large enough for the approximate shape of a patch to be deduced, and the motion of its 'centre' could also be determined. These observations have been described by Gerson (1950a, b). Finally, it has been shown that patches of sporadic-E ionization may be located at points some distance from a pulse transmitter. Scattered echoes are observed which are thought to be due to back scattering from the ground with intermediate reflection from patches of sporadic-E ionization at about 100 km height. If a directional aerial is used and is rotated, the echoes will appear and disappear. The directions of maximum amplitude give the directions of the patches, and the time delays of the echoes enable the distances to be determined. Villard, Peterson and Manning (1952) used a frequency of 14 Mc/s and the transmitter was arranged to transmit appreciable energy at low angles of elevation. The display took the form of a Plan Position Indicator of the type used in radar.

Unfortunately these methods have not so far yielded enough published results to justify a detailed comparison with the results of other methods. The movements reported by Gerson do, however, show that the centres of the patches tend to move around circles in a clockwise sense. This is the type of movement expected in the E region from the results of the measurements previously described (cf. fig. 19).

Findlay (1953) observed the phases of waves returned from isolated patches of ionization just below the E layer. The phases changed as they would if the patches were moving horizontally, not necessarily overhead. The directions of the movements could not be determined, but the velocities were between 20 m/sec and 150 m/sec. The velocity most frequently observed, viz. 70 m/sec, is the same as that obtained by the closely spaced receiver method (§2.2(*b*)). The sizes of the patches were of the order of a few hundred metres, which is much smaller than the sporadic-E patches observed by the other methods.

§7. Summary and Conclusions

The main results obtained by the various methods will now be summarized. Mention may be made here of a previous summary made in connection with a discussion before the Royal Astronomical Society (1951), and of summaries by Sheppard (1949) and Deb (1953). The general circulation in the upper atmosphere has been discussed by Kellogg and Schilling (1951).

For the purposes of the summary, the results will be divided into those which apply to the E region and to the F region.

7.1. *E Region* (*closely spaced receivers*; *meteor observations*)

The mean velocity of horizontal movements observed in the E region is about 80 m/sec. The velocity is made up of a systematic part, showing regular diurnal and seasonal variations, and an irregular part (closely spaced receivers, §2.2(*b*) and (*e*); meteor observations, §5.2). If the variations of the systematic part throughout the day are subjected to a Fourier analysis to determine the

semi-diurnal components, it is found that these can be well represented by a vector of constant magnitude of about 30 m/sec, which makes two complete revolutions every 24 hours. In the Northern hemisphere the rotation is clockwise and the phase is such that the movement is towards the North at 0300 and 1500 (closely spaced receivers, §2.2(*b*); meteor observations, §5.2). The rotating vector has smaller magnitude in summer than in winter and there is some evidence that the phase of its rotation is also subject to a seasonal variation (closely spaced receivers, §2.2(*b*)). In the Southern hemisphere the results are similar, except that the rotation is anticlockwise. The phase of the rotation is such that the movement is towards the South at 0530 and 1730 (mean of closely spaced receivers, §2.2(*b*), and meteor observations, §5.2). It would be expected that the movement would be towards the South in the Southern hemisphere when it is towards the North in the Northern hemisphere. There is thus a difference of 2 or 3 hours between the phases for the two hemispheres. The results for the Southern hemisphere are based on only a few months' data; this, together with the probable seasonal variation in the phase, could account for this difference.

Semi-diurnal variations in lunar time cannot be regarded as well established (§2.2(*b*)).

In addition to these movements, showing the diurnal variation, there is also a seasonal effect. Superimposed on the velocity, whose direction rotates twice in the 24 hours, there is, in the summer, a mean velocity of about 50 m/sec towards the East. Since this is greater than the variable component, the result is that the movement is towards the East for the greater part of the day in the summer, but that it changes direction during the day in the winter. At midday the resultant velocity is towards the East in summer, but towards the West in winter. This phenomenon is observed in both hemispheres (§2.2 (*b*)).

The velocity is sometimes observed to be different at heights differing by only a few kilometres (closely spaced receivers, §2.2(*e*); meteor observations, §5.2). It is possible that some of the random changes of velocity observed with the closely spaced receiver method may be due to small changes in the height of reflection of the waves, which are too small to be measurable with the techniques at present in use. There are also indications of a gradient of velocity with height of the order of 3·6 m/sec per km, with higher velocity at the greater height (meteor observations, §5.2).

Magnetic activity has little effect on the velocities observed in the E region, except during the largest magnetic disturbances.

7.2. *F Region* (*closely spaced receivers*; *movements of large irregularities*; *observations by radio astronomy*)

By day the EW component of velocity is consistently towards the East in the F region. The NS component shows a seasonal change, being towards the North in summer and towards the South in winter in the Northern hemisphere, and the opposite in the Southern hemisphere (closely spaced receivers, §2.2(*b*); movements of large irregularities, §3.2).

At night, the EW component is directed towards the West at all seasons and the NS component is small (closely spaced receivers, §2.2(*b*); movements of large irregularities, §3.2; radio astronomy, §4.2). The only exception to this comes from the observations by radio astronomy, which *sometimes* show

a reversal of the EW component near midnight, so that the velocity is towards the East after midnight. There is some evidence that this reversal is associated with the occurrence of magnetic activity, though the association is not well established. It should be remembered that the movements detected by the method of radio astronomy are at heights around 500 km, considerably higher than for the other methods.

The velocities in the F region are of the order of 100 m/sec under magnetically quiet conditions. During periods of magnetic activity the velocities are increased and, in general, a greater *K*-figure is associated with a greater velocity (closely spaced receivers, §2.2(*c*); movements of large irregularities, §3.2; radio astronomy, §4.2). This increase of velocity is more marked at the greater heights. A typical gradient of velocity with height is of the order of 1 m/sec per km.

It is not yet clear how the seasonal and diurnal changes of direction are modified during magnetically disturbed periods. The results from closely spaced receivers suggest that the movement is mainly towards the West during magnetic activity, with a small component towards the South, while the method of radio astronomy indicates a movement towards the West for the hours before midnight and towards the East for the hours after midnight. One method, dependent on observations at widely spaced points, suggests a movement towards the South during the daytime when magnetic activity is present (§3.2).

7.3. *Non-radio Methods*

Although our main intention was to survey the results of radio methods of observation, this report would be incomplete without some mention of the various non-radio methods which have been used to detect movements in the high atmosphere. We shall therefore briefly summarize these methods and the associated results.

(*a*) *Visual observations of meteor trails* (70–110 *km*).

The order of magnitude of the velocity of movement of visible meteor trails is 70 m/sec, and the directions are very variable. Some trails cover a range of heights, and are observed to have different directions of motion at levels only 10 km or less apart. The results obtained by this method have been summarized by Olivier (1942, 1947).

(*b*) *Observations of noctilucent clouds* (82 *km*).

Noctilucent clouds, seen mostly in polar regions, move horizontally with speeds of the order of 70 m/sec, and again the directions are not very consistent (Störmer 1935).

(*c*) *Self-luminous clouds* (120 *km*).

The self-luminous clouds observed by Hoffmeister (1946) are remarkable because they occur at heights well into the E region. The results for the magnitudes of the velocities were variable, but of the same order as those from other methods; they had, on the average, a greater velocity in winter than in summer. The directions of motion were predominantly towards the North-East in summer, while in winter there were two preferred directions, one just South of West and the other in the range North-East to North-West.

(*d*) *Observations of movements of the aurora* (100–300 *km*).

It has long been known that auroral rays often show an apparent movement towards the West, but this has been interpreted as a consequence of the Earth's rotation beneath an excitation pattern fixed with respect to the line joining the Earth to the Sun. However, some recent photographic observations made by Meinel and Schulte (1953), in which the EW movements were made more evident by an acceleration of the time scale, showed that this movement towards the West occurred only before midnight. After midnight movements towards the East were observed. It was concluded that the reversal could not be explained by the rotation of the Earth under a fixed excitation pattern, and must be evidence of mass movements of the atmosphere. The velocities observed by this method were in the range 300 m/sec to 1400 m/sec.

(*e*) *Movements of patches of emission in the light of the night sky.*

Studies of the forbidden radiation from atomic oxygen, of wavelength 5577 Å, by Roach, Williams, and Pettit (1953), showed that apparent movements of patches of increased emission could be detected. It was found that before midnight there was a movement towards the West. At midnight, when the intensity of the emission was a maximum, there was no movement. After midnight there was a small component of movement towards the East and a large component towards the South. The mean velocity observed was 150 m/sec. The height of the emission was not known with certainty, but was probably about 200 km.

It is interesting to note that the reversal of the EW component of movement at midnight, observed by the last two methods, agrees with that observed by the method of radio astronomy during periods of magnetic disturbance (§ 4.2).

(*f*) *Observations of the anomalous propagation of sound waves* (40 *km and* 90 *km*).

It is known that sound waves can be reflected down to the Earth from a region in the atmosphere where the temperature increases with height. There are two such regions, one near 40 km and the other near 90 km. Reflections from both levels have been observed. The wind velocity near the height of reflection can be inferred from observations of the direction in which anomalous propagation takes place. Thus Johnson and Hale (1953), observing in south-western United States, found that reflections from the lower level were obtained mainly to the east of a source of sound in winter and to the west of it in summer. This would indicate a mean velocity towards the East in winter and towards the West in summer, at 40 km. At 90 km the observations indicated an opposite movement—towards the West in winter and towards the East in summer.

ACKNOWLEDGMENTS

We are indebted to the following workers who have communicated results prior to publication:—

Mr. R. L. Adgie, Cavendish Laboratory, Cambridge.

Dr. W. J. G. Beynon and Mr. L. Thomas, University College of Swansea, University of Wales.

Mr. S. A. Bowhill, Cavendish Laboratory, Cambridge.

Mr. E. N. Bramley, Department of Scientific and Industrial Research, Radio Research Station, Slough.

Mr. G. J. Burtt, Dominion Physical Laboratory, Department of Scientific and Industrial Research, New Zealand.

Mr. W. G. Elford and Mr. D. S. Robertson, University of Adelaide, South Australia.

Mr. R. Greenstone, National Bureau of Standards, Washington, D.C., U.S.A.

Dr. L. Harang, Norwegian Defence Research Establishment.

Dr. A. Maxwell, Jodrell Bank Experimental Station, University of Manchester.

Dr. G. Newstead, working at Cavendish Laboratory, Cambridge, on leave from University of Tasmania.

Mr. B. W. Osborne, University of Malaya, Singapore.

One of us (M.S.) is indebted to the Department of Scientific and Industrial Research for the award of a maintenance grant.

References

Bowhill, S. A., 1953, *Thesis*, Cambridge University.

Booker, H. G., Ratcliffe, J. A., and Shinn, D. H., 1950, *Trans. Roy. Soc.* A, **242**, 579.

Bramley, E. N., 1953, *Proc. Roy. Soc.* A, **220**, 39.

Briggs, B. H., and Phillips, G. J., 1950, *Proc. Phys. Soc.* B, **63**, 907.

Briggs, B. H., Phillips, G. J., and Shinn, D. H., 1950, *Proc. Phys. Soc.* B, **63**, 106.

Chapman, J. H., 1953, *Canad. J. Phys.*, **31**, 120.

Deb, S., 1953, *J. Atmos. Terr. Phys.*, **4**, 28.

Elford, W. G., and Robertson, D. S., 1953, *J. Atmos. Terr. Phys.*, **4**, 271.

Fejer, J. A., 1953, *J. Atmos. Terr. Phys.*, **4**, 184.

Findlay, J. W., 1953, *J. Atmos. Terr. Phys.*, **3**, 73.

Gerson, N. C., 1950 a, *Nature, Lond.*, **166**, 316; 1950 b, *Geof. pura e appl.* (*Milan*), **18**, 3.

Greenhow, J. S., 1952, *J. Atmos. Terr. Phys.*, **2**, 282.

Hewish, A., 1951, *Proc. Roy. Soc.* A, **209**, 81; 1952, *Ibid.*, **214**, 494.

Hoffmeister, C., 1946, *Z. Meteorologie*, **1**, 33.

Johnson, C. T., and Hale, F. E., 1953, *J. Acoust. Soc. Amer.*, **25**, 642.

Kellogg, W. W., and Schilling, G. F., 1951, *J. Met.*, **8**, 222.

Krautkrämer, J., 1950, *Archiv Elekt. Übertragung*, **4**, 133.

Little, C. G., 1951, *Mon. Not. R. Astr. Soc.*, **111**, 289.

Little, C. G., and Lovell, A. C. B., 1950, *Nature, Lond.*, **165**, 422.

Little, C. G., and Maxwell, A., 1951, *Phil. Mag.*, **42**, 267; 1952, *J. Atmos. Terr. Phys.*, **2**, 356.

Manning, L. A., Villard, O. G., and Peterson, A. M., 1950, *Proc. Inst. Radio Engrs, N.Y.*, **38**, 877.

Maxwell, A., and Little, C. G., 1952, *Nature, Lond.*, **169**, 746.

Meinel, A. B., and Schulte, D. H., 1953, *Astrophys. J.*, **117**, 454.

Millman, G. H., 1952, *Ann. Geophys.*, **8**, 365.

Mitra, S. N., 1949, *Proc. Instn Elect. Engrs*, Pt. III, **96**, 441.

Munro, G. H., 1948, *Nature, Lond.*, **162**, 886; 1950, *Proc. Roy. Soc.* A, **202**, 208; 1953, *Nature, Lond.*, **171**, 693.

Olivier, C. P., 1942, *Proc. Amer. Phil. Soc.*, **85**, 93; 1947, *Ibid.*, **91**, 315.

Pawsey, J. L., 1935, *Proc. Camb. Phil. Soc.*, **31**, 125.

Phillips, G. J., 1951, *Proc. Instn Elect. Engrs*, Pt. III, **98**, 237; 1952, *J. Atmos. Terr. Phys.*, **2**, 141.

Price, R. E., 1953, *Nature, Lond.*, **172**, 115.

Roach, F. E., Williams, D. R., and Pettit, H., 1953, *Astrophys. J.*, **117**, 456.

Robertson, D. S., Liddy, D. T., and Elford, W. G., 1953, *J. Atmos. Terr. Phys.*, **4**, 255.

Royal Astronomical Society Geophysical Discussion, 1951, *Nature, Lond.*, **167**, 626.

Ryle, M., and Hewish, A., 1950, *Mon. Not. R. Astr. Soc.*, **110**, 381.

Salzberg, C. D., and Greenstone, R., 1951, *J. Geophys. Res.*, **56**, 521.

Sheppard, P. A., 1949, *Science Progress*, **37**, 488.

Smith, F. G., 1950, *Nature, Lond.*, **165**, 422.

Störmer, C., 1935, *Astrophys. Norv.*, **1**, No. 3.

Villard, O. G., Peterson, A. M., and Manning, L. A., 1952, *Proc. Inst. Radio Engrs, N.Y.*, **40**, 992.

Wilkes, M. V., 1949, *Oscillations of the Earth's Atmosphere* (Cambridge: University Press).